

EV 357836964 US

(19) RÉPUBLIQUE FRANÇAISE  
INSTITUT NATIONAL  
DE LA PROPRIÉTÉ INDUSTRIELLE  
PARIS

(11) N° de publication : 2 748 479  
(à n'utiliser que pour les  
commandes de reproduction)

(21) N° d'enregistrement national : 96 05854

(51) Int Cl<sup>®</sup> : C 07 K 14/33, C 12 N 15/56, 15/70, 15/74, 15/79, 1/21,  
C 12 P 21/02, G 01 N 33/68 // (C 12 N 15/56, C 12 R 1:145)

(12) DEMANDE DE BREVET D'INVENTION

A1

(22) Date de dépôt : 10.05.96.

(30) Priorité :

(43) Date de la mise à disposition du public de la  
demande : 14.11.97 Bulletin 97/46.

(56) Liste des documents cités dans le rapport de  
recherche préliminaire : *Se reporter à la fin du  
présent fascicule.*

(60) Références à d'autres documents nationaux  
apparentés :

(71) Demandeur(s) : INSTITUT PASTEUR — FR.

(72) Inventeur(s) : LEIBOVITZ EMMANUELLE et BEGUIN  
PIERRE.

(73) Titulaire(s) : .

(74) Mandataire : REGIMBEAU.

(54) POLYPEPTIDE COMPORTANT UN NOUVEAU DOMAINE COHÉSINE DE TYPE II, COMPOSITION  
ENZYMATIQUE EN COMPORTANT ET FRAGMENTS D'ADN CODANT POUR CES POLYPEPTIDES.

(57) La présente invention concerne notamment un poly-  
peptide comprenant un domaine cohésine de type II, ca-  
ractérisé en ce qu'il est capable de se fixer à un domaine  
dockerine de type II, d'une protéine de charpente, d'un  
complexe enzymatique, d'une bactérie cellulolytique, no-  
tamment de la protéine CipA de *Clostridium thermocellum*.

La présente invention fournit également une nouvelle  
protéine SdbA ("Scaffoldin dockerin binding protein") de  
*Clostridium thermocellum*.

La présente invention fournit une composition enzymati-  
que comprenant plusieurs enzymes réunis par l'intermé-  
diaire de molécules d'association comprenant des domai-  
nes cohésine et des domaines dockerine.

FR 2 748 479 - A1



POLYPEPTIDE COMPORTANT UN NOUVEAU DOMAINE COHÉSINE DE  
TYPE II, COMPOSITION ENZYMATIQUE EN COMPORTANT ET  
FRAGMENTS D'ADN CODANT POUR CES POLYPEPTIDES

5           La présente invention concerne des domaines protéiques  
susceptibles d'interagir de façon non covalente et permettant d'agencer en  
complexes multiprotéiques définis des polypeptides d'intérêt biochimique  
ou biologique pour les faire agir ensemble simultanément ou de manière  
séquentielle afin de potentialiser leur synergie. Elle concerne également  
10 les fragments d'ADN codant pour lesdits fragments protéiques. La présente  
invention concerne enfin des compositions enzymatiques permettant  
d'associer plusieurs enzymes pour les faire agir ensemble simultanément ou  
de manière séquentielle afin de potentialiser leur synergie. Par exemple,  
dans le cas d'une action séquentielle, plusieurs types d'enzymes à activité  
15 différente peuvent agir successivement sur un même mélange de substrats.

Les cellulases de plusieurs bactéries cellulolytiques sont  
organisées en complexe enzymatique comportant des sous-unités à activité  
catalytique interagissant avec un polypeptide sans activité catalytique  
appelé "protéine de charpente". Cette interaction se réalise via des  
20 domaines des sous-unités à activité catalytique appelés "domaines  
dockerine" et des domaines répétés de la protéine de charpente appelés  
"domaines cohésine" de taille plus importante que les domaines dockerine  
des unités catalytiques.

A ce jour, seuls les domaines cohésine des protéines de  
25 charpente ont été identifiés, ces domaines sont appelés dans la présente  
description, domaines cohésine de type I.

En particulier, Clostridium thermocellum, une bactérie Gram  
positive, thermophile et anaérobie, produit un complexe cellulolytique à  
masse moléculaire élevée dénommé cellulosome (15, 16, 21). Ce complexe est  
30 initialement fixé à la surface cellulaire et est ensuite libéré dans le milieu.  
Le cellulosome est composé d'au moins 15 polypeptides différents,  
comprenant de nombreuses  $\beta$ -1,4-endoglucanases, au moins une  
cellobiohydrolase (23) et plusieurs hémicellulases ( $\beta$ -1,4-xylanases,  
lichénases) (22). Les composants catalytiques sont liés de manière non  
35 covalente à une sous-unité de charpente non catalytique, dénommée CipA  
(pour Cellulosome Integrating Protein) (37).

La protéine CipA et des composants similaires identifiés dans les complexes cellulolytiques d'autres *Clostridium* cellulolytiques sont des protéines de charpente ou "scaffoldines" (2).

Le mode de fixation des sous-unités catalytiques à la protéine  
5 Cip A a été élucidé (références n° 8, 33). Chaque sous-unité catalytique contient un segment dupliqué et conservé de 23 résidus, constituant un domaine dockerine (2). Les domaines dockerine entrent en interaction avec un ensemble de domaines de liaison complémentaires, ou domaines cohésine (2).

10 Ces domaines, dont neuf copies sont présentes dans la séquence de CipA, sont très semblables entre eux, particulièrement les domaines 4 à 8, qui possèdent plus de 95 % de résidus identiques (11).

Il a été montré que l'on peut greffer un domaine dockerine sur  
une protéine ne faisant pas partie du cellulosome, par exemple  
15 l'endoglucanase CelC de *C. thermocellum*, et que celle-ci acquiert de ce fait la capacité de se fixer sur CipA (32).

Cette observation a suggéré la possibilité d'utiliser l'affinité  
entre domaines cohésine et domaines dockerine afin de créer des complexes  
artificiels incorporant diverses protéines fusionnées à des domaines  
20 dockerine adéquats, interagissant avec les domaines cohésine de la protéine de charpente (2, 32). De tels complexes pourraient trouver diverses applications biotechnologiques. En modifiant de manière contrôlée, la composition de cellulosomes naturels, il pourrait être possible d'optimiser leur activité vis-à-vis de substrats cellulotiques définis. On peut également  
25 envisager d'améliorer le processus de dégradation d'autres substrats complexes et insolubles, faisant appel à des enzymes de spécificité complémentaire, et dont l'action synergique serait potentialisée par une association en complexes multienzymatiques. De même, l'association physique d'enzymes effectuant des réactions séquentielles permet  
30 d'accélérer celles-ci lorsque la vitesse de diffusion du produit de la première réaction vers le deuxième site réactionnel est limitante (L. Bülow et K. Mosbach, Multienzyme systems obtained by gene fusion, Trends in Biotechnol. 9, 226-231). Par ailleurs, l'utilité de complexes multiprotéiques n'est pas limitée à l'association d'enzymes. La construction de complexes  
35 protéiques multifonctionnels est en effet susceptible de donner lieu à une grande variété d'applications, discutées dans la référence (2).

Cependant, la construction de complexes de stoechiométrie et de topologie définies se heurte à une difficulté importante. Tous les domaines cohésine connus jusqu'à présent sont très semblables quant à leur séquence et à leur spécificité de liaison. Par exemple, il a été montré que

5 CelS, une des sous-unités catalytiques du cellulosome, peut se lier de façon équivalente aux domaines cohésine (18b) 1, 2, et 9 de CipA, et vraisemblablement à tous les autres domaines cohésine de celle-ci. En conséquence, il n'est pas possible de programmer la liaison d'une protéine de fusion, porteuse d'un domaine dockerine, à un domaine cohésine défini

10 de la protéine de charpente.

Les domaines cohésine connus jusqu'à ce jour, possédant une forte similitude de séquence et de spécificité de liaison, ont été groupés sous le nom de domaines cohésine de type I. De même, les domaines dockerine portés par les sous-unités catalytiques du cellulosome, et capables de se lier

15 aux domaines cohésine de type I, sont appelés domaines dockerine de type I.

Il existe cependant à l'extrémité COOH-terminale de CipA un domaine présentant une similitude de séquence éloignée avec les domaines dockerine de type I, mais incapable de se lier aux domaines cohésine de type I. Il permet aux protéines qui le portent de se fixer à trois polypeptides

20 exocellulaires de *C. thermocellum*. La structure et la fonction de ces polypeptides sont inconnues (29).

L'invention repose sur la caractérisation d'un gène, *sdbA* ("scaffolding dockerin binding protein"), qui a été cloné et séquencé, dont le produit SdbA est capable de se fixer spécifiquement au domaine COOH-

25 terminal de CipA, à l'exclusion des domaines dockerine de type I portés par les sous-unités catalytiques du cellulosome. La caractérisation du polypeptide SdbA montre qu'il comporte une région spécifique responsable de la liaison avec le domaine COOH-terminal de CipA, et dont la séquence est très différente de celle des domaines cohésine de type I. Cette région, ainsi

30 que les segments polypeptidiques de séquence et de spécificité d'interaction similaire, sont nouveaux et appelés domaines cohésine de type II. De même, la région COOH-terminale de CipA est appelée domaine dockerine de type II. L'utilisation de domaines cohésine et dockerine de type II, éventuellement en conjonction avec des domaines cohésine et dockerine de type différent

35 (par exemple de type I) permet de construire des complexes protéiques mieux définis.

L'intérêt des domaines cohésine de type II selon la présente invention est de présenter une spécificité de reconnaissance différente de celle des domaines cohésine de protéine de charpente connus précédemment, notamment ceux de la protéine CipA.

5 La présente invention concerne plus particulièrement des domaines cohésine de type II ainsi que des domaines dockerine de type II.

La présente invention concerne notamment des composés sur lesquels sont capables de se fixer de façon covalente ou non au moins un domaine cohésine de type II ou un domaine dockerine de type II.

10 Plus particulièrement ces composés sont des peptides, polypeptides ou protéines, mais il peut s'agir de lipides ou de glycosides ou bien de molécules de type mixte telles que protéoglycane, lipopolysaccharide par exemple. Il est possible de prévoir d'autres types de molécules notamment des marqueurs ou par exemple des molécules  
15 chimiques thérapeutiques ou non.

Un domaine cohésine de type II est un domaine protéique qui se lie de façon spécifique avec le domaine dockerine de CipA correspondant au domaine dockerine de type II tel qu'il sera défini ci-après. De préférence l'affinité du complexe ainsi formé sera au moins de  $10^5$  M/L tel que mesuré  
20 par la méthode décrite dans SALAMITOU et al (réf. 28).

La séquence du domaine dockerine de CipA est celle correspondant à l'IDS n° 4.

Un domaine cohésine de type II peut correspondre à des séquences naturelles, il peut notamment s'agir de domaine provenant de  
25 bactéries cellulolytiques notamment des Clostridium comme cela sera décrit ci-après pour SdbA.

Mais de tels domaines sont également présents sur les protéines OlpB et ORF2p.

La notion de domaine cohésine de type II incorpore également  
30 des séquences protéiques non naturelles pour autant qu'elles puissent se lier avec le domaine dockerine de type II de CipA.

Il peut alors s'agir notamment de domaines homologues aux domaines naturels ou de fragments de ces domaines mais il est possible de prévoir également des domaines entièrement synthétiques obtenus par  
35 exemple en utilisant certains acides aminés non naturels, ou bien en utilisant des éléments améliorant l'affinité.

Par "protéine homologue" ou "séquence homologue" on entend selon la présente invention toute protéine, polypeptide ou peptide présentant une homologie de séquence d'au moins 25 % par rapport au domaine cohésine de type II notamment celui correspondant à SdbA, ladite  
5 séquence conservant les propriétés de liaison spécifique au domaine dockerine concerné, notamment au domaine dockerine de CipA.

Par "fragment de protéines" ou "fragment de séquences" on entend un fragment d'au moins 50 acides aminés conservant les propriétés de liaison spécifique au domaine dockerine concerné, notamment au  
10 domaine dockerine de CipA.

Il faut rappeler qu'un domaine cohésine de type II doit présenter une bonne affinité pour le domaine dockérine correspondant mais ne doit présenter que pas ou peu d'affinité pour le domaine dockérine de type différent notamment de type I.

15 La présente invention concerne également les composés comportant un domaine dockérine de type II, c'est à dire un domaine protéique qui se lie de façon spécifique avec un domaine cohésine de type II et ce avec une affinité d'au moins  $10^5$  M/L mesuré comme précédemment.

Cette définition n'est pas redondante, en effet il faut bien  
20 comprendre qu'à partir du domaine dockérine de type II de CipA il est possible de définir un certain nombre de domaines cohésine de type II, lesquels peuvent permettre de définir de nouveaux domaines dockerine de type II lesquels comme précédemment peuvent être d'origine naturelle, mais peuvent être constitués de fragments de domaines de séquences  
25 homologues ou bien éventuellement comme cela a été indiqué précédemment, comporter des séquences entièrement synthétiques avec éventuellement des acides aminés non naturels.

La liaison entre un domaine cohésine et un domaine dockerine de type II sera dénommée ci-après par simplification interaction C/D de type II, le complexe ainsi formé étant dénommé soit complexe C/D de type II lorsqu'il ne comporte qu'une seule interaction C/D de type II soit complexe multimérique lorsqu'il comporte au moins une interaction C/D autre que de type II, interaction C/D de type I par exemple et/ou d'autres formes d'interactions : avidine/biotine, antigène/anticorps par exemple. De  
30 préférence, les complexes multimérique selon l'invention comportent essentiellement des interactions de type C/D.  
35

Ainsi la présente invention, grâce à ces différentes interactions, permet de cibler l'intégration dans un complexe d'enzymes différentes et de fournir un complexe artificiel utilisant notamment une protéine de charpente comportant des domaines dockerine de spécificité de liaisons différentes afin d'agencer de manière spécifique diverses protéines porteuses de domaines cohésine correspondants.

Plus particulièrement, la présente invention fournit un polypeptide ayant un domaine cohésine de type II, selon l'invention, caractérisé en ce qu'il est capable de se fixer au domaine dockerine COOH-terminal de la protéine de charpente CipA du complexe cellulolytique de Clostridium thermocellum. Toute protéine ou peptide présentant ou comportant une séquence ayant plus de 25 % de résidus d'acides aminés identiques avec un domaine cohésine de type II de SdbA entre dans la définition de l'invention. En particulier, il s'agit d'une protéine de Clostridium thermocellum ou d'un fragment de celle-ci.

Dans un mode plus particulier de réalisation, la présente invention a pour objet une protéine SdbA ("scaffolding dockerin binding protein") de Clostridium thermocellum de poids moléculaire apparent de 68kDa ( $\pm 10$  %) comportant un domaine cohésine qui est capable de se fixer avec un domaine dockerine de type II notamment de la protéine CipA de Clostridium thermocellum.

Le polypeptide SdbA du complexe cellulolytique de Clostridium thermocellum, a une séquence de 631 acides aminés substantiellement telle que représentée sur l'IDS n°1.

La présente invention a permis d'identifier le domaine de la protéine SdbA capables de se fixer au domaine dockerine de CipA. En particulier, le domaine cohésine comprend une séquence de la région N-terminale de la protéine de 184 acides aminés substantiellement telle que représentée dans l'IDS n° 1 de l'acide aminé n° 27 à l'acide aminé n° 210 de la séquence de la protéine ou une séquence homologue ou un fragment de cette séquence ou d'une séquence homologue capable de se fixer à un domaine dockerine de la protéine CipA, par exemple, un fragment de ces séquences d'au moins 50 acides aminés capable de se fixer à un domaine dockerine de la protéine CipA.

La présente invention a permis d'identifier des domaines cohésine de type II d'autres protéines de Clostridium thermocellum, en particulier des protéines OlpB et ORF2p (9, 17). SdbA présente une homologie de séquence avec les séquences répétées N-terminale de OlpB et  
5 ORF2p.

Le segment polypeptidique comprenant les résidus 26-199 de la protéine OlpB, qui présente une forte similitude de séquence avec les résidus 27-191 de SdbA, peut également fixer le domaine C-terminal de CipA

Ainsi outre des fragments de la protéine SdbA ou d'une  
10 protéine homologue interagissant avec un domaine dockerine d'une protéine de charpente selon l'invention, la présente invention a donc également pour objet des fragments de OlpB et ORF2p, de séquences similaires au domaine cohésine de type II de SdbA.

La présente invention a donc pour objet tout polypeptide  
15 comprenant comme domaine cohésine la séquence correspondant substantiellement à l'une des séquences de la protéine OlpB choisies parmi la séquence des acides aminés n° 28 au n° 192, la séquence des acides aminés n° 207 au n° 363, la séquence des acides aminés n° 409 au n° 565 et la séquence des acides aminés n° 607 au n° 763 de l'IDS n° 2 ou une séquence  
20 homologue à l'une de ces séquences ou un fragment de ces séquences d'au moins 50 acides aminés, capable de se fixer à un domaine dockerine de la protéine CipA.

La présente invention a également pour objet tout polypeptide comprenant un domaine cohésine qui a substantiellement pour séquence en  
25 acides aminés, une séquence de la protéine ORF2p choisie parmi la séquence des acides aminés n° 38 à 195 et la séquence des acides aminés n° 209 à 365 de l'IDS n° 3, ou une séquence homologue à ces séquences ou un fragment de ces séquences d'au moins 50 acides aminés capable de se fixer à un domaine dockerine de la protéine CipA.

30 La présente invention a également pour objet tout polypeptide qui comporte un segment de séquence de plus de 50 acides aminés, présentant plus de 25 % de résidus identiques avec l'un des segments de l'IDS n° 1, de l'IDS n° 2, ou de l'IDS n° 3 décrits ci-dessus, et capable de fixer le domaine dockerine de CipA.



Parmi les composés selon la présente invention comportant un domaine cohésine de type II ou dockerine de type II il faut citer par exemple les enzymes, les récepteurs, les antigènes, les anticorps ou un de leurs fragments comportant entre 20 et 100 acides aminés.

5 Dans le cas particulier où la protéine est une enzyme, il s'agira par exemple d'une cellulase permettant une meilleure hydrolyse de substrat cellulosique ou tout autre type d'enzyme hydrolytique.

Dans le cas où le composé selon l'invention est essentiellement une protéine, les parties de la protéine peuvent être fusionnées audit domaine cohésine ou dockerine par l'intermédiaire d'un  
10 fragment polypeptidique. La liaison peut être également une liaison non covalente, par exemple une liaison conformationnelle.

La présente invention a également pour objet un fragment d'ADN codant pour un composé selon l'invention lorsque celui-ci est un  
15 polypeptide ou codant pour la protéine SdbA ou un fragment de celle-ci, lorsque le composé selon l'invention comporte d'autres éléments que le polypeptide ou la protéine, l'invention concerne également le fragment d'ADN codant pour le polypeptide ou la protéine.

La présente invention repose en partie sur le clonage  
20 moléculaire et le séquençage du gène dénommé sdbA, dont le produit se fixe spécifiquement au domaine dockerine porté par CipA. Des segments du gène ont été sous-clonés et exprimés séparément, pour permettre d'identifier la région du polypeptide responsable de la fixation du domaine dockerine de CipA. Il s'agit du fragment d'ADN comprenant un cadre de lecture ouvert  
25 de 1 893 nucléotides, et codant pour le polypeptide de 631 aminoacides dénommé SdbA, ayant une masse moléculaire calculée de 68 577 Da.

La présente invention a donc également pour objet un fragment d'ADN représenté substantiellement par la séquence 1 à 1893 de l'IDS n° 1 codant pour la protéine SdbA ainsi qu'une souche de E. coli  
30 déposée à la CNCM de l'Institut Pasteur sous le n° I-1684 transformée par le plasmide PCT1830 comportant un fragment d'ADN correspondant à cette séquence codant pour la protéine SdbA.

La présente invention a en outre pour objet un fragment d'ADN qui a pour séquence essentiellement les nucléotides 82 à 573 dans  
35 l'IDS n° 1 codant pour le domaine cohésine de la protéine SdbA ainsi qu'une souche de E. coli déposée à la CNCM de l'Institut Pasteur sous le n° I-1683 transformée par le plasmide pCT1801 comportant un fragment d'ADN correspondant à cette séquence de 1893 pb.

De même la présente invention a également pour objet un fragment d'ADN caractérisé en ce qu'il a substantiellement pour séquence l'une des séquences codant pour un domaine cohésine de la protéine OlpB choisies parmi la séquence des nucléotides 85 à 570, la séquence des  
5 nucléotides 619 à 1095, la séquence des nucléotides 1225 à 1689 et la séquence des nucléotides 1819 à 2189 dans l'IDS n° 2 ainsi qu'un fragment d'ADN caractérisé en ce qu'il a substantiellement pour séquence l'une des séquences codant pour un domaine cohésine de ORF2 choisies parmi la séquence des nucléotides 109 à 582 et la séquence des nucléotides n° 625 à  
10 1092 dans l'IDS n° 3.

La présente invention a également pour objet des fragments d'ADN qui ont pour une séquence une séquence complémentaire ou homologue ou complémentaire de l'homologue d'un des fragments d'ADN tels que définis ci-dessus.

15 Par "fragment d'ADN homologue" on entend des fragments qui codent pour des polypeptides homologues comme cela a été décrit précédemment.

La présente invention a également pour objet des fragments d'ADN capables de s'hybrider dans des conditions faiblement stringentes  
20 (19) avec un fragment d'ADN selon l'invention tel que défini précédemment.

La présente invention concerne également des complexes comportant au moins un composé tel que décrit précédemment lié par une interaction C/D de type II avec un composé comportant au moins un  
25 domaine dockerine de type II, chaque composé constituant un élément du complexe.

Il s'agit notamment d'un complexe multimérique caractérisé en ce que au moins deux des "éléments" du complexe sont liés par une interaction C/D de type II. De préférence le complexe comportera au moins  
30 trois "éléments" dont deux des "éléments" sont liés par une interaction autre que C/D de type II par exemple par une interaction C/D de type I.

Par "élément" on désignera :  
un composé selon l'invention qui pourra éventuellement comporter un autre domaine de liaison : interaction C/D de type I par exemple, ou bien  
35 un composé comportant un seul domaine de liaison différent de l'interaction C/D de type II mais capable de se fixer sur un composé selon l'invention.

En utilisant judicieusement les divers types d'interactions, il est possible d'obtenir des complexes ayant des structures variées. La structure du complexe multimérique ou la structure du complexe de type II selon l'invention peuvent être ainsi de type linéaire ou greffée ou bien de type mixte.

Un complexe multimérique de type linéaire comprend un enchaînement de composés selon l'invention, ne comportant que deux domaines de liaison chacun. Un tel complexe est représenté à la figure 1 B.

Au contraire, une structure greffée comporte en général une molécule de structure avec un certain nombre de domaines de liaison et des greffons protéines par exemple, ne comportant qu'un seul domaine, ce type de structure est schématisé à la figure 1A et 1C.

Bien entendu, il est possible de prévoir des structures qui combinent ces deux structures de base, on peut même prévoir des structures cycliques.

En effectuant une fixation de façon séquentielle on peut ainsi obtenir un complexe de structure bien défini, ce qui est particulièrement intéressant pour obtenir des complexes enzymatiques.

Les composés selon la présente invention peuvent être obtenus par génie génétique lorsqu'il s'agit de protéines. Lorsque les composés comportent des éléments non protéiques, ceux-ci peuvent être greffés par des moyens connus notamment par réactions chimiques pour les liaisons covalentes ou par des liaisons non covalentes.

Une première façon de mettre en oeuvre l'invention consiste à fusionner au moyen du génie génétique des domaines cohésine respectivement dockerines de type différent, de façon à construire des protéines de charpente comportant ces domaines en nombre et en ordre définis. Parallèlement, des domaines dockerines respectivement cohésines adéquats sont greffés sur des protéines étrangères, par exemple des enzymes, que l'on désire associer dans un ordre choisi le long de la protéine de charpente, on obtient ainsi une structure greffée correspondant à la figure 1A ou C. Ce type de mise en oeuvre conduit à des complexes se rapprochant du cellulosome naturel.

Ces complexes selon l'invention pourront comprendre de préférence pour chaque composé des segments peptidiques de jonction, de longueur et de séquence appropriées. Par exemple, les constructions

reprendront les segments de jonction naturels riches en proline et/ou hydroxy amino acides présents dans les polypeptides naturels. L'incorporation des protéines que l'on désire associer s'effectue par l'intermédiaire d'un domaine cohésine ou dockerine greffé, par exemple au  
5 moyen du génie génétique.

Dans le complexe selon la présente invention, le nombre d'éléments du multimère est compris entre 1 et 50 éléments associés entre eux et de préférence 1 et 20.

Dans un mode de réalisation, chaque élément du complexe  
10 comprend des domaines cohésines ou des domaines dockerine.

Mais il est possible de prévoir des éléments comportant des domaines cohésines et dockerines.

La présente invention a également pour objet un fragment d'ADN codant pour un élément du complexe selon l'invention.

D'une manière générale, la présente invention a également  
15 pour objet les vecteurs d'expression comprenant un fragment d'ADN selon l'invention placé sous le contrôle d'éléments assurant son expression dans une cellule hôte de type eucaryote ou dans un hôte bactérien tel qu'une souche de E.coli transformée par un vecteur d'expression selon l'invention,  
20 et un procédé de préparation d'un polypeptide selon l'invention ou d'une protéine selon l'invention caractérisé en ce qu'on réalise la culture de cellules hôtes transformées à l'aide d'un vecteur d'expression selon l'invention ou par culture d'une souche de E.coli selon l'invention.

Enfin la présente invention fournit une composition  
25 enzymatique comprenant plusieurs enzymes réunis afin de les faire agir ensemble et le cas échéant potentialiser leur synergie, par l'intermédiaire d'un complexe multimérique sur chacun desquels est couplé une enzyme différente.

La présente invention concerne des compositions comportant  
30 au moins un complexe multimérique présentant au moins un domaine d'interaction C/D de type II.

En particulier, une composition enzymatique selon l'invention peut comprendre deux enzymes réunies afin de les faire agir ensemble et le cas échéant potentialiser leur synergie, par l'intermédiaire d'un complexe  
35 selon l'invention comportant une première enzyme, et une seconde enzyme liée par interaction C/D de type II.

Dans une variante avantageuse de réalisation, ledit complexe comporte un polypeptide comprenant un domaine cohésine selon l'invention, couplé à un domaine dockerine de la protéine Cip A couplé à un premier enzyme, et un deuxième élément comprend un domaine dockerine d'une sous unité catalytique du complexe cellulolytique de Clostridium thermocellum couplé à une seconde enzyme, qui se liera au domaine cohésine.

Les complexes multimériques selon l'invention sont plus particulièrement utilisables lorsque lesdits complexes multimériques potentialisent la synergie des éléments des complexes, notamment lorsqu'il s'agit d'une composition enzymatique.

La présente invention concerne également un procédé de détection d'un antigène ou d'un anticorps par la mise en contact d'un complexe multimérique selon l'invention avec une solution contenant un anticorps ou un antigène d'intérêt et la révélation de la réaction entre le complexe multimérique et l'antigène ou l'anticorps.

La révélation peut se faire par marquage radioactif du complexe anticorps ou antigène ou par visualisation en utilisant des marquages non isotopiques, par exemple de type avidine - biotine ou tout autre marquage équivalent.

D'autres caractéristiques et avantages de la présente invention apparaîtront à la lumière de la description détaillée qui va suivre. Cette description fait référence aux figures 1 à 6.

La figure 1 schématise la structure de complexe multimérique selon l'invention.

La figure 2 représente une carte de restriction de la région comprenant le gène *sdbA*, et construction de pCT1830, pCT1831 et pCT1832, codant pour SdbA-N, SdbA-C et SdbA, respectivement. E: EcoRI; K: KpnI; P: PstI; Sa: Sall; Sc: SacI; Sp: SphI; SCM: site de clonage multiple. Les positions des segments codant pour les diverses régions identifiées dans SdbA sont indiquées par des cadres de dessins différents. Les nombres se réfèrent à la séquence nucléotidique (figure 3). Les nucléotides qui ont été changés dans la séquence amplifiée par PCR sont indiqués en gras. L'ADN du vecteur PQE-30 est indiqué par un trait mince. La séquence de pQE-30, codant pour 6 résidus histidine, est représentée par un cadre qui n'est pas à l'échelle. La transcription de *sdbA* va de gauche à droite.

La figure 3 représente une séquence nucléotidique de la région codant pour le gène *sdbA*. Le site de liaison ribosomique supposé est souligné. Les diverses régions identifiées dans *SdbA* sont indiquées par des cadres de même dessin que sur la figure 2. SLR: site de liaison ribosomique.

5 La figure 4 représente l'alignement du domaine cohésine de *SdbA* et des domaines cohésine de *OlpB* et *ORF2p* (9). Les résidus qui sont identiques ou similaires à la majorité des séquences représentées sont indiqués sur un fond ombré. La numérotation des résidus commence avec des codons d'initiation supposés. Les aminoacides similaires sont: F, I, V, L et  
10 M; R et K; S et T; D et E; N et Q; et F, Y et W.

La figure 5 représente la similarité entre les résidus 264 à 275 de *SdbA* et d'un motif présent dans les protéines M de *Streptococcus pyogenes*. M1: (numéro de dépôt GenBank x72752), M9 (24), PAM (3), M12 (26). Pour chaque protéine, la numérotation commence avec le codon  
15 d'initiation supposé. Les résidus qui sont identiques ou similaires dans la majorité des séquences représentées sont indiqués sur un fond ombré. Les critères de similarité sont les mêmes que pour la figure 4.

La figure 6 représente l'alignement des segments répétés COOH-terminaux de *SdbA* avec les séquences similaires d'autres protéines de  
20 surface cellulaire. *OlpA*: protéine A de couche externe de *C. thermocellum* (9); *OlpB*: protéine B de couche externe de *C. thermocellum* (9); *Pul*: pullulanase de *T. thermosulfurigenes* EM1 (20); *Bsph*: protéine de couche S de *B. sphaericus* (4). Pour chaque protéine, la numérotation commence au niveau du codon d'initiation supposé. Les résidus qui sont semblables ou identiques  
25 dans au moins huit segments sont indiqués sur fond ombré. Les critères de similarité sont les mêmes que pour la figure 4.

## I. MATERIEL ET METHODES

### 30 1. Souches bactériennes, plasmides et conditions de culture

Les souches bactériennes et les plasmides utilisés dans cette étude sont récapitulés dans le tableau 1. La souche TG1 d'*Escherichia coli* a été utilisée pour le clonage et le séquençage. Les protéines ont été produites dans *E. coli* M15 (pREP4).

35 *C. thermocellum* a été cultivé dans des conditions anaérobies, à 60°C dans du milieu CM3-3 complété avec 5 g de cellobiose par litre (31).

On a cultivé *E. coli* à 37°C, dans du milieu de Luria Bertani (19). On a ajouté des antibiotiques en fonction des plasmides présents dans l'hôte: 100 µg/ml de ticarcilline, 30 µg/ml de chloramphénicol, 25 µg/ml de kanamycine.

## 5 2. Manipulations d'ADN

L'ADN génomique de *C. thermocellum* a été purifié par la méthode de Marmur modifiée par Quiviger et coll. (25). D'autres manipulations d'ADN ont été effectuées selon Ausubel et coll. (1). On a utilisé les enzymes de restriction en suivant les recommandations des fournisseurs.

10 Les amorces oligonucléotidiques ont été synthétisées par Eurogentec SA (Sering, Belgique) ou Genset SA (Paris, France). On a effectué l'amplification par PCR selon Saiki et coll. (27), en utilisant 100 pmoles de chaque amorce oligonucléotidique dans un mélange réactionnel de 100 µl. MgCl<sub>2</sub> a été ajouté jusqu'à une concentration finale de 2 mM. On a effectué  
15 35 cycles d'amplification. Les paramètres étaient les suivants: hybridation: 1 minute à 65°C; extension: 1 minute à 72°C; et dénaturation: 1 minute à 94°C. On a toujours vérifié la séquence des fragments clonés obtenus par PCR.

## 3. Construction de la banque génomique de *C. thermocellum*

L'ADN de *C. thermocellum* a été partiellement digéré par  
20 Sau3AI, et les fragments ont été séparés sur un gradient de saccharose. Des fragments de plus de 12 kb ont été insérés dans le plasmide pUC18 coupé par BamHI, et traités par de la phosphatase alcaline bactérienne (Ready-to-go, Pharmacia). Des cellules de *E. coli* TG1 ont été transformées par électroporation et étalées en présence de 0,8 mg de  
25 5-bromo-4-chloro-3-indolyl-β-D-galactoside par plaque et 0,2 mg d'isopropyl-β-D-thiogalactoside (IPTG) par plaque.

## 4. Criblage de colonie et repérage de protéines transférées sur membrane

On a criblé comme décrit (8) les clones recombinants, en recherchant la fixation de CelC-DsCipA marquée au <sup>125</sup>I.

30 Pour identifier les polypeptides porteurs de domaines cohésine de type II, on a analysé les protéines par SDS-PAGE (14) et on les a transférées sur une membrane en Nylon (Hybond-N+, Amersham) (1). La membrane a été mise à incuber avec CelC-DsCelD et CelC-DsCipA marquées au <sup>125</sup>I, lavée et autoradiographiée comme décrit précédemment (29, 32).

## 35 5. Séquences d'ADN et analyse des séquences

Les fragments de restriction appropriés de pCT1801 ont été sous-clonés dans le plasmide pBCSK, et on a engendré des délétions emboîtées en utilisant de l'exonucléase III et de la nucléase S1 (nécessaire Erase-a-base, Promega), comme indiqué par le fournisseur. On a séquencé les matrices monocaténares conformément à la méthode de terminaison de chaîne didésoxy de Sanger et coll. (30), en utilisant les nécessaires Sequenase et Taquence (USB-Amersham). La séquence a été déterminée au moins une fois sur chaque brin. L'analyse par ordinateur des données des séquences a été effectuée au moyen du logiciel Sequence Analysis Software Package de Genetic Computer Group, version 7 (University of Wisconsin) (6).

#### 6. Construction de clones d'expression et purifications de protéine

En utilisant le vecteur pQE-30, on a construit des clones produisant en excès des formes de SdbA intactes ou comprenant des délétions. Ainsi, on a fait fusionner la séquence codant pour le polypeptide recherché avec un segment codant pour 6 résidus His, pour faciliter la purification (13). Afin de cloner le fragment codant pour le domaine NH<sub>2</sub>-terminal de SdbA, on a synthétisé par PCR un fragment de 670 pb encadré par BamHI et PstI (figure 1). L'amorce directe était

5'-CTG CCG GCG GGA TCC GCA AGG GCA GAT-3'

et l'amorce inverse était

5'-ACT TTT GCA GAA TTT TCT GCA GGC G-3'.

Le fragment a été inséré entre les sites BamHI et PstI de pQE30, pour donner pCT1830. Le polypeptide codé par pCT1830 a été dénommé SdbA-N.

Pour cloner la région codant pour les domaines COOHterminaux de SdbA, on a fait digérer par BamHI le plasmide pCT1801. Les extrémités ont été complétées et converties en extrémités franches à l'aide du fragment de Klenow de l'ADN polymérase. Après nouvelle coupure par PstI, le fragment de 1,4 kb, codant pour les domaines COOH-terminaux, a été purifié et inséré dans le vecteur pQE-30 qui avait été digéré par HindIII, traité par le fragment de Klenow de l'ADN polymérase et digéré à nouveau par PstI. Le plasmide résultant a été dénommé pCT1831, et le polypeptide codé a été dénommé SdbA-C.

Le plasmide pCT1832, exprimant la séquence complète de SdbA, a été construit par insertion du fragment BamH-PstI de 670 pb (voir plus haut) dans le plasmide pCT1831 digéré par BamHI et PstI.



La production et la purification des protéines ont été effectuées au moyen du système Qjaexpress (QJAGEN Inc.). Des cultures de 1 litre ont été mises à incuber à 37°C jusqu'à une DO<sub>600</sub> de 0,7. On y a ensuite ajouté de l'IPTG jusqu'à une concentration finale de 0,3 mM, et les cultures  
5 ont été mises à nouveau à incuber pendant 5 heures à 37°C. On a remis les cellules en suspension dans 80 ml de Tris.HCl 50 mM, pH 7,5 (tampon A) et on les a lysées au moyen d'une presse de French Aminco, sous une pression de 100 MPa. On a centrifugé l'extrait à 9 000 g pendant 20 minutes afin d'éliminer les débris cellulaires. On a injecté le surnageant dans une  
10 colonne de 8 ml de résine Ni-NTA équilibrée avec du tampon A, on a lavé la colonne avec du tampon A et on l'a éluée avec le même tampon contenant 250 mM d'imidazole. Les fractions éluées ont été dialysées pendant une nuit à 4°C, contre 1 litre de tampon A. Les protéines purifiées ont été conservées à -80°C.

15 7. Détermination de séquence amino-terminale d'acides aminés

50 pmoles de chaque polypeptide à séquencer ont été séparées par SDS-PAGE et transférées pendant une nuit, à la température ambiante, à 850 mA sur une membrane en PVDF [poly(chlorure de vinylidène)] hydrophobe (Problott, Applied Biosystem) traitée par du méthanol à 100 %, au moyen d'un système Trans-Blot Cell (BioRad) contenant 50 mM de Tris  
20 (base), 50 mM de tampon acide borique. On a coloré les bandes au noir amide à 0,003 %, on les a excisées, et on a déterminé la séquence amino-terminale des polypeptides par la méthode d'Edman, en utilisant un appareil de séquençage 473A ou Procise HT (Applied Biosystem).

25

## II. RESULTATS

1. Clonage d'un gène codant pour un polypeptide se fixant spécifiquement au domaine dockerine de CipA

30 On a criblé 1 600 clones recombinants en recherchant la fixation de CelC-DsCipA marquée au <sup>125</sup>I. Huit clones indépendants ont été marqués spécifiquement. Les contrôles effectués avec de la CelC-DsCelD marquée au <sup>125</sup>I ont indiqué que la fixation était spécifique pour le domaine dockerine de CipA (figure 2).

35 Tous les segments clonés s'hybrident avec la même région du génome de C. thermocellum (données non représentées), dont la carte est représentée sur la figure 1. Ces cartes de restriction sont en accord avec les

fragments de restriction révélés par analyse Southern blot dans l'ADN de C. thermocellum (données non représentées). Les segments ne s'hybridaient pas et n'ont pas de fragments de restriction en commun avec la région comprenant cipA et olpA (9). Dans la région couverte par les fragments  
5 clonés, un segment de 1,6 kb, compris entre le site PstI et la limite gauche de l'insert porté par pCT1801 (figure 1), est nécessaire et suffisant pour coder pour un polypeptide capable de fixer le domaine dockerine de CipA. Le gène correspondant a été dénommé sdbA.

## 2. Analyse de la séquence

10 La séquence du gène de la SdbA est représentée sur la figure 3. La séquence codante comprend 1 893 nucléotides. Le codon d'initiation ATG est précédé d'un site de liaison ribosomique supposé. Le polypeptide codé, composé de 631 aminoacides, a une masse moléculaire calculée de 68 577 Da. La structure de domaines de la protéine est représentée sur les figures 1 et 3.  
15 Un peptide signal supposé de 26 résidus aminoacide est localisé à l'extrémité NH<sub>2</sub>-terminale du polypeptide (36). Des alignements avec d'autres protéines indiquent la présence de trois régions distinctes dans SdbA. La région N-terminale, composée de 156 résidus aminoacide, est semblable aux segments N-terminaux répétés de OlpB (dénommé précédemment ORF1p) et  
20 ORF2p de C. thermocellum, deux polypeptides dont les gènes sont localisés immédiatement en aval de cipA (9) (figure 4). Un espaceur de 56 résidus, riche en Pro/Thr/Ser, sépare cette région du reste de la protéine. La région centrale est composée de 215 aminoacides, avec de nombreux résidus Lys. Cette région comprend une courte séquence d'acides aminés semblable à un  
25 segment présent dans les protéines M de Streptococcus pyogenes (figure 5). La région COOH-terminale est composée de ces segments répétés qui sont très semblables aux segments dénommés SLH (S-layer homologous = homologues à la couche S), présents dans plusieurs protéines localisées sur la surface cellulaire de diverses bactéries (9, 18) (figure 6).

## 30 3. Identification du domaine responsable de la fixation du domaine dockerine de CipA

Afin d'identifier le domaine responsable de la fixation du domaine dockerine de CipA, on a comparé les propriétés de liaison de polypeptides dérivés de SdbA. Le gène sdbA et des sous-fragments  
35 appropriés ont été fusionnés avec le vecteur d'expression pQE-30 codant pour His<sub>6</sub>, et les polypeptides correspondants ont été purifiés par chromatographie d'affinité à Ni (24). Les masses moléculaires apparentes de

la protéine SdbA intacte et du fragment contenant les régions centrale et C-terminale sont de 60 kDa et 36 kDa, respectivement, en accord avec les masses prédites à partir de la séquence (figure 7A). La masse moléculaire apparente du domaine NH<sub>2</sub>-terminal était égale à 35 kDa, et était supérieure à la masse moléculaire calculée à partir de la séquence (22 715 Da). Toutefois, le fragment comprend le segment de jonction riche en résidus Pro, ce qui peut expliquer une lente migration dans la SDS-PAGE (10). Les préparations de SdbA intacte et du polypeptide COOH-terminal contenaient l'une et l'autre un second polypeptide de 24 kDa. Dans les deux cas, la séquence NH<sub>2</sub>-terminale de ce polypeptide est SKYAVSY, ce qui indique qu'elle est dérivée de la région COOH-terminale contenant les segments SLH répétés de SdbA. Etant donné que les segments SLH répétés ne contiennent pas de groupement de résidus histidine, le fragment COOH-terminal est probablement lié aux polypeptides intacts. En effet, il a été rapporté que des polypeptides contenant des segments SLH répétés s'auto-associent (17).

L'analyse du criblage de colonies, en utilisant comme sonde CelC-DsCipA marquée au <sup>125</sup>I, a confirmé que le produit du gène sdbA se fixait au domaine dockerine de CipA (figure 7B). La fixation au fragment NH<sub>2</sub>-terminal est moins intense, mais décelable. On n'a pas pu déceler de fixation au fragment C-terminal. Etant donné que la région NH<sub>2</sub>-terminale de SdbA est semblable aux segments NH<sub>2</sub>-terminaux répétés d'OlpB, on a contrôlé si CelC-DsCipA se fixait à MalE-ORFlp-N, une protéine chimère comprenant le premier segment NH<sub>2</sub>-terminal répété d'OlpB fusionné à la protéine de fixation du maltose, MalE (17). La colonne 5 de la figure 7B indique que MalE-ORFlp-N a été marquée. Aucune fixation n'a été observée avec MalE-ORFlp-C, qui consiste en les segments SLH C-terminaux d'OlpB fusionnés à MalE. Ni SdbA, ni ORFlp-N, ni ORFlp-C n'ont été marquées après incubation avec CelC-DsCelD marquée au <sup>125</sup>I (données non représentées).

Des protéines portant des domaines dockerine peuvent être marquées au <sup>125</sup>I et utilisées comme sondes pour la détection de protéines contenant des domaines cohésine complémentaires (29, 32). Ainsi, on peut isoler des clones exprimant des polypeptides contenant des domaines cohésine, et on peut identifier les domaines cohésine (8). Dans la présente invention, on a appliqué la même stratégie pour cloner le gène sdbA et pour identifier le domaine cohésine responsable de la fixation du domaine dockerine de CipA. On a obtenu un seul gène. Il se peut que d'autres gènes

codant pour des protéines ayant des propriétés similaires aient échappé à la détection, en raison d'une absence d'expression spontanée.

Sur les trois polypeptides, pl70, pll6 et p60, qui se sont précédemment révélés fixer le domaine dockerine de CipA (29), pl70 et pll6  
5 sont trop longs pour être codés par sdbA, même en tenant compte de modifications post-traductionnelles, telles qu'une glycosylation. Le polypeptide p60 se révèle le seul candidat possible.

La figure 7 indique que le domaine cohésine se trouve dans la région NH<sub>2</sub>-proximale de SdbA. Le signal détecté avec le fragment  
10 NH<sub>2</sub>-terminal est plus faible qu'avec la protéine entière; toutefois, on n'a pu détecter aucun signal en utilisant une quantité semblable du fragment COOH-terminal. Le fait de tronquer SdbA peut avoir affecté l'affinité ou la stabilité du polypeptide NH<sub>2</sub>-terminal résiduel. Ou encore, la fixation à la  
15 nitrocellulose peut altérer la conformation du domaine cohésine, tandis que la fixation de la protéine intacte à la membrane peut être médiée par des régions du polypeptide non requises pour la fixation de la sonde marquée.

Contrairement au domaine dockerine de CipA, qui est clairement apparenté aux domaines dockerine présents dans les sous-unités catalytiques, le domaine cohésine de SdbA ne présente pas de similarité  
20 évidente avec les domaines cohésine de CipA et OlpA. Toutefois, il est semblable aux segments répétés localisés à l'extrémité NH<sub>2</sub>-terminale de OlpB et ORF2p (9). En effet, CelC-DsCipA marquée au <sup>125</sup>I se fixe spécifiquement au premier segment répété NH<sub>2</sub>-terminal d'OlpB. Ainsi, les  
25 domaines NH<sub>2</sub>-terminaux de SdbA, OlpB et très probablement ORF2p représentent un nouveau type de domaine cohésine. C'est pourquoi, selon la présente invention on les dénomme "domaines cohésine de type II", et "domaines cohésine de type I" les domaines cohésine rencontrés dans CipA et à l'extrémité NH<sub>2</sub>-terminale d'OlpA.

Les trois protéines OlpB, ORF2p et SdbA, qui sont connues  
30 comme contenant les domaines cohésine de type II, portent également des segments répétés SLH. Dans tous les cas étudiés jusqu'à présent, les segments répétés SLH se rencontrent dans des protéines qui sont associées à la surface cellulaire de bactéries, et des preuves biochimiques indiquent qu'ils se  
35 fixent à des composants de l'enveloppe cellulaire (17). Ainsi, SdbA peut être localisée sur la surface cellulaire, au même titre qu'OlpA (28) et OlpB (17). La similarité entre la région centrale de SdbA et une région présente dans les protéines M de Streptococcus vient à l'appui de cette hypothèse. Il a été

supposé que dans les protéines M, cette région peut entrer en interaction avec des glucides de la paroi cellulaire (34). Prises dans leur ensemble, ces considérations suggèrent que SdbA, OlpB et éventuellement ORF2p sont des composants de l'enveloppe cellulaire qui sont impliqués dans la fixation de  
5 cellulosomes à la surface cellulaire.

Alors que SdbA ne porte qu'un seul domaine cohésine, ces domaines sont répétés deux fois dans ORF2p et quatre fois dans OlpB. Ainsi, jusqu'à quatre molécules de CipA portant des sous-unités catalytiques fixées pourraient être groupées autour d'une molécule d'OlpB. Toutefois, ce fait  
10 seul ne suffit pas pour rendre compte de la formation d'agrégats très volumineux (polycellulosomes) allant jusqu'à 80 MDa, comme rapporté dans la référence (5). De tels agrégats doivent impliquer d'autres interactions, éventuellement au niveau des segments répétés SLH, qui sont reconnus se  
lier entre eux (17).

15

**TABLEAU 1**  
**Souches bactériennes et plasmides**

5	Souches et plasmides	Caractères significatifs	Source de Référence
<b>Souches</b>			
10	<u>Escherichia coli</u>	TG1 [ $\Delta$ (lac-pro) thi supE hsdD5/ F tra-36proA+B+lacI <sub>q</sub> lacZ $\Delta$ M15] M15 (pREP4)	(12)  (7,35), nécessaire QJAexpress® QJAGEN Inc.
15	<u>Clostridium thermocellum</u> NCIB 10682		
20	Plasmides		
	pUC18		(38)
	pBCSK-		Stratagene®
	pQE-30		nécessaire
			QJAexpress®
25			QJAGEN Inc.
	pCT1801	dérivé de pUC18 contenant un fragment Sau3A codant pour SdbA	n° CNCM I-1684
30	pCT1830	dérivé de pQE-30 codant pour le domaine cohésine de SdbA soudé à 6 résidus His	n° CNCM I-1684
	pCT1831	dérivé de pQE-30 codant pour les régions centrale et COOH- terminale de SdbA soudées à 6 résidus His	la présente étude
35	pCT1832	dérivé de pQE-30 codant pour SdbA soudé à 6 résidus His	"

## BIBLIOGRAPHIE

1. F. M. AUSUBEL, R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN, J. A. SMITH, and K. STRUHL, 1990, Current Protocols in Molecular Biology, Greene Publishing and Wiley Interscience, New York.  
5
2. E.A. BAYER, E. MORAG, and R. LAMED, 1994, The cellulosome - a treasuretrove for biotechnology. Trends in Biotechnol. 12:379-386.
- 10 3. A. BERGE and U. SJOBRING, 1993, PAM, A novel plasminogen-binding protein from *Streptococcus pyogenes*. J. Biol. Chem. 268:25417-24.
- 15 4. R.D. BOWDITCH, P. BAUMANN, and A.A. YOUSTEN, 1989, Cloning and sequencing of the gene encoding a 125-kilodalton surface-layer protein from *Bacillus sphaericus* 2362 and a related cryptic gene. J. Bacteriol. 171:4178-4188.
- 20 5. COUGHLAN, M.P., K. HON-NAMI, H. HON-NAMI, L.G. LJUNGDAHL, J. J. PAULIN, and W.E. RIGSBY, 1985, The cellulolytic enzyme complex of Clostridium thermocellum is very large, Biochem. Biophys. Res. Commun. 130:904-909.
- 25 6. DEVEREUX, J., P. HAEBERLI, and O. SMITHIES, 1984, A comprehensive set of sequence analysis programs for the VAS. Nucleic Acids Res. 12:387-395.
7. FARABAUGH, P.J., 1978, Sequence of the *lacI* gene, Nature 274:765-769.
- 30 8. FUJINO, T., P. BEGUIN, and J.P. AUBERT, 1992, Cloning of a Clostridium thermocellum DNA fragment encoding polypeptides that bind the catalytic components of the cellulosome, FEMS Microbiol. Lett. 94:165-170.
- 35 9. FUJINO, T., P. BEGUIN and J.P. AUBERT, 1993, Organization of a Clostridium thermocellum gene cluster encoding the cellulosomal scaffolding protein CipA and a protein possibly involved in the attachment of the cellulosome to the cell surface. J. Bacteriol. 175:1891-1899.

10. FURTHMAYR, H., and R. TIMPL, 1971, Characterization of collagen peptides by sodium dodecylsulfate-polyacrylamide electrophoresis. *Anal. Biochem.* 41:510-516.
- 5 11. GERNGROSS, U.T., M.P.M. ROMANIEC, N.S. HUSKISSON, and A.L. DEMAINE, 1993, Sequencing of a Clostridium thermocellum gene (CipA) encoding the cellulosomal S<sub>L</sub>-protein reveals an unusual degree of internal homology. *Mol. Microbiol.* 8:325-334.
- 10 12. GIBSON, T.J. 1984, Studies on the Epstein-Barr virus genome, 1984, University of Cambridge, Cambridge, UK.
- 15 13. JANKNECHT, R., G. DE MARTYNOFF, J. LOU, R. A. HIPSKIND, A. NORDHEIM, and G.G. STUNNENBERG, 1991, Rapid and efficient purification of native histidine-tagged protein expressed by recombinant vaccinia virus. *Proc. Natl. Acad. Sci. USA* 88:8972-8976.
- 20 14. LAEMMLI, U.K., 1970, Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature* 227:680-685.
- 25 15. LAMED R., R. KENIG, E. SETTER, and E.A. BAYER, 1985, Major characteristics of the cellulolytic system of Clostridium thermocellum coincide with those of the purified cellulosome, *Enzyme Microb. Technol.* 7:37-41.
- 30 16. LAMED R., E. SETTER, R. KENIG, and E.A. BAYER, 1983, The cellulosome : a discrete cell surface organelle of Clostridium thermocellum with exhibits separate antigenic, cellulose-binding and various cellulolytic activities. *Biotechnol. Bioeng. Symp.* 13:163-181.
17. LEMAIRE M., H. OHAYON, P. GOUNON, T. FUJINO and P. BEGUIN, 1995, OlpB, a new outer layer protein of Clostridium thermocellum, and binding of its S-layer-like domain to components of the cell envelope, *J. Bacteriol.* 77:2451-2459.



18. LUPAS A., H. ENGELHARDT, J. PETERS, U. SANTARIUS, S. VOLKER, and W. BAUMEISTER, 1994, Domain structure of the *Acetogenium kivui* surface layer revealed by electron crystallography and sequence analysis, J. Bacteriol. 176:1224-1233.
- 5 18b. LYTLE B., C. HYERS, K. KRUUS and J. H. WU, 1996, Interactions of the Cels' binding ligand with various receptor domains of the clostridium the mocoellum cellulosomal scaffolding protein, CipA, 25 J. Bacteriol. 178: 1200-1203.
- 10 19. MANIATIS T., E.F. FRITSCH, and J. SAMBROOK, 1982, Molecular Cloning, a Laboratory Manual. p. In Editor (ed.) Vol. Cold Spring Harbor Laboratory, N.Y.
- 15 20. MATUSCHEK M., G. BURCHHARDT, K. SAHM, and H. BAHL, 1994, Pullulanase of *Thermoanaerobacter thermosulfurigenes* EM1 (*Clostridium thermosulfurogenes*), molecular analysis of the gene, composite structure of the enzyme, and a common model for its attachment to the cell surface, J. Bacteriol. 176:3295-3302.
- 20 21. McBEE R.H., 1948, The culture and physiology of a thermophilic cellulose-frementing bacterium, J. Bacteriol. 56:653-663.
- 25 22. MORAG E., E. A. BAYER, and R. LAMED, 1990, Relationship of cellulosomal and non-cellulosomal xylanases of Clostridium thermocellum to cellulose-degrading enzymes, J. Bacteriol. 172:6098-6105.
- 30 23. MORAG E., I. HALEVY, E.A. BAYER, and R. LAMED, 1991, Isolation and properties of a major cellobiohydrolase from the cellulosome of Clostridium thermocellum, J. Bacteriol., 173:4155-4162.
- 35 24. PODBIELSKI A., J. HAWLITZKY, T.D. PACK, A. FLOSDORFF, and M.D. BOYLE, 1994, A group A streptococcal Enn protein potentially resulting from intergenomic recombination exhibits atypical immunoglobulin-binding characteristics, Mol. Microbiol. 12:725-736.

25. QUIVIGER B., C. FRANCHE, G. LUTFALLA, R.D., R. HASELKORN, and C. ELMERICH, 1982, Cloning of a nitrogen fixation (*nif*) gene cluster of *Azospirillum brasilense*, *Biochimie* 64:495-502.
- 5 26. ROBBINS J.C., J.G. SPANIER, S.J. JONES, W.J. SIMPSON, and P.P. CLEARY, 1987, *Streptococcus pyogenes* type 12 M protein gene regulation by upstream sequences, *J. Bacteriol.* 169:5633-5640.
- 10 27. SAIKI R.K., D.H. GELFAND, S. STOFFEL, S.J. SCHARF, R. HIGUCHI, G.T. HORN, K.B. MULLIS and H.A. ERLICH, 1988, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science* 239:487-491.
- 15 28. SALAMITOU S., M. LEMAIRE, T. FUJINO, H. OHAYON, P. GOUNON, P. BEGUIN, and J.P. AUBERT, 1994, Subcellular localization of *Clostridium thermocellum* ORF3p, a protein carrying a receptor for the docking sequence borne by the catalytic components of the cellulosome, *J. Bacteriol.* 176:2828-2834.
- 20 29. SALAMITOU S., O. RAYNAUD, M. LEMAIRE, M. COUGHLAN, P. BEGUIN and J.P. AUBERT, 1994, Recognition specificity of the duplicated segments present in *Clostridium thermocellum* endoglucanase CelD and in the cellulosome-integrating protein CipA, *J. Bacteriol.* 176:2822-2827.
- 25 30. SANGER F., S. NICKLEN, and A.R. COULSON, 1977, DNA sequencing with chain-terminating inhibitors, *Proc. Natl. Acad. Sci. USA* 74:5463-5467.
- 30 31. TAILLIEZ P., H. GIRARD, J. MILLET, and P. BEGUIN, 1989, Enhanced cellulose fermentation by an asporogenous and ethanol-tolerant mutant of *Clostridium thermocellum*, *Appl. Environ. Microbiol.* 55:207-211.
32. TOKATLIDIS K., P. DHURJATI, and P. BEGUIN, 1993, Properties conferred on *Clostridium thermocellum* endoglucanase CelC by grafting the duplicated segment of endoglucanase CelD. *Protein Engng* 6:947-952.
- 35 33. TOKATLIDIS K., S. SALAMITOU, P. BEGUIN, P. DHURJATI, and J.P. AUBERT, 1991, Interaction of the duplicated segment carried by *Clostridium thermocellum* cellulases with cellulosome components, *FEBS Lett.* 291:185-188.

34. VIJAYKUMAR P. and V.A. FISCHETTI, 1988, Isolation and characterization of the cell-associated region of group A streptococcal M6 proteins, J. Bacteriol. 170:
- 5 35. VILLAREJO M.R., and I. ZABIN, 1974,  $\beta$ -galactosidase from termination and deletion mutant strains, J. Bacteriol. 120:466-474.
36. VON HEIJNE G., 1983, Patterns of amino acids near signal-sequence cleavage sites, Eur. J. Biochem. 133:17-21.
- 10 37. WU J.H.D. and A.L. DEMAINE, 1988, Proteins of the Clostridium thermocellum cellulase complex responsible for degradation of crystalline cellulose, p. 117-131, In J.P. AUBERT, P. BEGUIN, and J. MILLET (ed.), FEMS Symposium n° 43 Biochemistry and Genetics of Cellulose Degradation, Academic Press., London & New York.
- 15 38. YANISCH-PERRON C., J. VIEIRA, and J. MESSING, 1985, Improved M13 phage cloning vectors and host strains, nucleotide sequences of the M13mp18 and pUC19 vectors, Gene 33:103-119.

(1) INFORMATIONS GENERALES:

(A) NOM: INSTITUT PASTEUR  
(B) RUE: 28 Rue du Docteur Roux  
(C) VILLE: PARIS  
(E) PAYS: FRANCE  
(F) CODE POSTAL: 75724 CEDEX 15

(ii) TITRE DE L' INVENTION: "POLYPEPTIDE COMPORTANT UN DOMAINE COHESINE, COMPOSITION ENZYMATIQUE EN COMPORTANT ET FRAGMENTS D'ADN CODANT POUR CES POLYPEPTIDES"

(iii) NOMBRE DE SEQUENCES: 4

(iv) FORME DECHIFFRABLE PAR ORDINATEUR:

(A) TYPE DE SUPPORT: Floppy disk  
(B) ORDINATEUR: IBM PC compatible  
(C) SYSTEME D' EXPLOITATION: PC-DOS/MS-DOS  
(D) LOGICIEL: PatentIn Release #1.0, Version #1.30 (OEB)

(2) INFORMATION FOR LA SEQ ID NO: 1:

(i) CARACTERISTIQUES DE LA SEQUENCE:

(A) LONGUEUR: 1893 paires de bases  
(B) TYPE: nucléotide  
(C) NOMBRE DE BRINS: simple

(ii) TYPE DE MOLECULE: ADN

(ix) CARACTERISTIQUE:

(A) NOM/CLE: SdbA de Clostridium thermocellum  
(B) EMPLACEMENT: 1..1893

(xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 1:

ATG AGG AAG AAA AAA AGA TTA ATA TCA TTA CTG CTT GCG GTT TTT ATC 48  
Met Arg Lys Lys Lys Arg Leu Ile Ser Leu Leu Leu Ala Val Phe Ile  
1 5 10 15

GCC GTT GCA TGT CTG CCG GCG GGA ATT GCA AGG GCA GAT AAA GCC TCG 96  
Ala Val Ala Cys Leu Pro Ala Gly Ile Ala Arg Ala Asp Lys Ala Ser  
20 25 30

AGC ATT GAG CTT AAG TTT GAC CGC AAT AAG GGA GAA GTT GGA GAT ATA Ser Ile Glu Leu Lys Phe Asp Arg Asn Lys Gly Glu Val Gly Asp Ile 35 40 45	144
CTT ATT GGT ACC GTA AGG ATA AAC AAT ATC AAG AAT TTC GCA GGA TTT Leu Ile Gly Thr Val Arg Ile Asn Asn Ile Lys Asn Phe Ala Gly Phe 50 55 60	192
CAG GTA AAC ATT GTA TAT GAT CCA AAA GTC TTA ATG GCT GTT GAC CCT Gln Val Asn Ile Val Tyr Asp Pro Lys Val Leu Met Ala Val Asp Pro 65 70 75 80	240
GAA ACG GGG AAA GAA TTT ACT TCT TCA ACA TTT CCG CCA GGA CGC ACT Glu Thr Gly Lys Glu Phe Thr Ser Ser Thr Phe Pro Pro Gly Arg Thr 85 90 95	288
GTA CTG AAA AAC AAT GCT TAC GGC CCA ATA CAG ATT GCG GAC AAT GAT Val Leu Lys Asn Asn Ala Tyr Gly Pro Ile Gln Ile Ala Asp Asn Asp 100 105 110	336
CCG GAA AAA GGG ATA CTG AAC TTC GCG CTT GCA TAT TCA TAT ATT GCG Pro Glu Lys Gly Ile Leu Asn Phe Ala Leu Ala Tyr Ser Tyr Ile Ala 115 120 125	384
GGA TAC AAA GAA ACA GGA GTA GCG GAG GAA AGC GGC ATA ATT GCG AAA Gly Tyr Lys Glu Thr Gly Val Ala Glu Glu Ser Gly Ile Ile Ala Lys 130 135 140	432
ATT GGA TTT AAA ATA CTC CAG AAA AAG AGC ACT GCC GTA AAA TTC CAG Ile Gly Phe Lys Ile Leu Gln Lys Lys Ser Thr Ala Val Lys Phe Gln 145 150 155 160	480
GAT ACA TTA AGC ATG CCC GGA GCT ATT TCG GGA ACA CAG CTG TTT GAC Asp Thr Leu Ser Met Pro Gly Ala Ile Ser Gly Thr Gln Leu Phe Asp 165 170 175	528
TGG GAC GGA GAA GTT ATT ACC GGA TAT GAG GTA ATA CAG CCG GAT GTG Trp Asp Gly Glu Val Ile Thr Gly Tyr Glu Val Ile Gln Pro Asp Val 180 185 190	576
CTG AGT TTG GGT GAC GAG CCT TAT GAG ACA CCG GGA ACG GAT ATT CCG Leu Ser Leu Gly Asp Glu Pro Tyr Glu Thr Pro Gly Thr Asp Ile Pro 195 200 205	624
ATA TCC GAC AAT CCG GCA GCA ACT CCG TCA TCC ACG CCG TCA GTT ACT Ile Ser Asp Asn Pro Ala Ala Thr Pro Ser Ser Thr Pro Ser Val Thr 210 215 220	672
CCT TCA CCG GAA GTT AAA CCG ACT CAG ACG CCT TCG CCT GCA GAA AAT Pro Ser Pro Glu Val Lys Pro Thr Gln Thr Pro Ser Pro Ala Glu Asn 225 230 235 240	720

TCT GCA AAA GTG GAG CTT GAA CCT GTG TTG GAT AAT GCA ACA GGA GAA Ser Ala Lys Val Glu Leu Glu Pro Val Leu Asp Asn Ala Thr Gly Glu 245 250 255	768
GCA AAG GCG GCA ATA GAT GAA GAA AAA TTA AAC AAG GCT CTT GAT GAA Ala Lys Ala Ala Ile Asp Glu Glu Lys Leu Asn Lys Ala Leu Asp Glu 260 265 270	816
GCG AAA AAA TCG GAA GAT GAC AAA CTT GTG GAA CTT AAC ATA AAG AAG Ala Lys Lys Ser Glu Asp Asp Lys Leu Val Glu Leu Asn Ile Lys Lys 275 280 285	864
GTT GAA AAT GCC GAT GCT TAC ATA CAA CAG CTT CCG GCG AAA TTC CTG Val Glu Asn Ala Asp Ala Tyr Ile Gln Gln Leu Pro Ala Lys Phe Leu 290 295 300	912
ATA AAA AGT GAC GCC GAA TAT AAG CTG AGA ATA GCT ACA GAG CAG GGA Ile Lys Ser Asp Ala Glu Tyr Lys Leu Arg Ile Ala Thr Glu Gln Gly 305 310 315 320	960
ATT ATA GAA GTA CCG GCC AAC ATG CTG AAT ACT GCG GAT ATT TCA AAG Ile Ile Glu Val Pro Ala Asn Met Leu Asn Thr Ala Asp Ile Ser Lys 325 330 335	1008
CTT GTA AAA AAT GAC TCC GTT GTT GAA TTC GTC ATA AGA AAA GTA AAA Leu Val Lys Asn Asp Ser Val Val Glu Phe Val Ile Arg Lys Val Lys 340 345 350	1056
GTC GAT GAA CTT GGT GCA GAG CTC AAA GAG AAG ATA GGC AAC AGG CCG Val Asp Glu Leu Gly Ala Glu Leu Lys Glu Lys Ile Gly Asn Arg Pro 355 360 365	1104
GTG ATT GAC ATA AGC GTG GTT GTT GAC GGC AAA AAA GTT GAA TGG AGC Val Ile Asp Ile Ser Val Val Val Asp Gly Lys Lys Val Glu Trp Ser 370 375 380	1152
AAT TAC AAA GCC AAG GTT AAA ATA TCA ATT CCT TAC AAG CCT GAT GCA Asn Tyr Lys Ala Lys Val Lys Ile Ser Ile Pro Tyr Lys Pro Asp Ala 385 390 395 400	1200
AAA GAG CTG GAG AAC CAC GAG CAT ATT GTT GTA CTC CAT ATT GAT GAC Lys Glu Leu Glu Asn His Glu His Ile Val Val Leu His Ile Asp Asp 405 410 415	1248
GCC GGC AAG GCA GTT TCC GTA CCC AGC GGA AAA TAT GAA CCT TCT TTG Ala Gly Lys Ala Val Ser Val Pro Ser Gly Lys Tyr Glu Pro Ser Leu 420 425 430	1296
GGC GTC GTT ACG TTT GAG ACG AAT CAT TTA AGC AAG TAT GCG GTT TCA Gly Val Val Thr Phe Glu Thr Asn His Leu Ser Lys Tyr Ala Val Ser 435 440 445	1344

TAT GTT TAC AAG ACT TTC GCG GAT ATT GGT TCA TAT GCC TGG GCT AAA Tyr Val Tyr Lys Thr Phe Ala Asp Ile Gly Ser Tyr Ala Trp Ala Lys 450 455 460	1392
AAG CAG ATA GAG GTT TTG GCT TCC AAA GGA GTA ATT AAC GGT ACA TCC Lys Gln Ile Glu Val Leu Ala Ser Lys Gly Val Ile Asn Gly Thr Ser 465 470 475 480	1440
GAT ACC ACT TTT ACG CCC CAG GCA GAC ATA ACA AGG GCG GAT TTC ATG Asp Thr Thr Phe Thr Pro Gln Ala Asp Ile Thr Arg Ala Asp Phe Met 485 490 495	1488
ATA CTT CTT GTA AAG GCA CTG GGA TTG ACT GCC GAG GTT ACT TCC AAT Ile Leu Leu Val Lys Ala Leu Gly Leu Thr Ala Glu Val Thr Ser Asn 500 505 510	1536
TTT GAT GAT GTG TCC GAA AAA GAC TAC TAT TAT GAA TAC GTG GGA ATT Phe Asp Asp Val Ser Glu Lys Asp Tyr Tyr Tyr Glu Tyr Val Gly Ile 515 520 525	1584
GCA AAA GAG CTT GGA ATT ACG ACA GGA GTC GGA AAC AAC AAG TTC AAT Ala Lys Glu Leu Gly Ile Thr Thr Gly Val Gly Asn Asn Lys Phe Asn 530 535 540	1632
CCG AAA GCC AAA ATT ACA AGA CAG GAT ATG ATG GTA CTT ACA ACA AAT Pro Lys Ala Lys Ile Thr Arg Gln Asp Met Met Val Leu Thr Thr Asn 545 550 555 560	1680
GCT CTC AGG ATT GCA GGA AAA ATA TCG AGC ACA GGA ACC CGC GCT GAT Ala Leu Arg Ile Ala Gly Lys Ile Ser Ser Thr Gly Thr Arg Ala Asp 565 570 575	1728
GTT GAA AGA TTT TCG GAC AAG GAC CAG ATA GCT TCA TAT GCG GTT GAA Val Glu Arg Phe Ser Asp Lys Asp Gln Ile Ala Ser Tyr Ala Val Glu 580 585 590	1776
GGC GTT GCA ACC TTG GTA AAA GAA GGT ATT GTA GTG GGA AGC GGC GAT Gly Val Ala Thr Leu Val Lys Glu Gly Ile Val Val Gly Ser Gly Asp 595 600 605	1824
ATT ATA AAT CCA AGG GGA AAT GCT TCA AGA GCC GAA CTT GCA GCA ATC Ile Ile Asn Pro Arg Gly Asn Ala Ser Arg Ala Glu Leu Ala Ala Ile 610 615 620	1872
ATA TAC AAG ATT TAC TAC AAG Ile Tyr Lys Ile Tyr Tyr Lys 625 630	1893

## (3) INFORMATIONS POUR LA SEQ ID NO: 2:

## (i) CARACTERISTIQUES DE LA SEQUENCE:

(A) LONGUEUR: 4992 paires de bases

(B) TYPE: nucléotide

(C) NOMBRE DE BRINS: simple

## (ii) TYPE DE MOLECULE: ADN

## (ix) CARACTERISTIQUE:

(A) NOM/CLE: OlpB de Clostridium thermocellum

(B) EMPLACEMENT:1..4992

## (xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 2:

ATG AAA CGA AAA AAT AAA GTA TTA TCA ATT TTG TTA ACT CTG CTG CTA	48
Met Lys Arg Lys Asn Lys Val Leu Ser Ile Leu Leu Thr Leu Leu Leu	
1 5 10 15	
ATA ATC TCT ACC ACA TCC GTA AAC ATG TCT TTT GCT GAA GCA ACT CCA	96
Ile Ile Ser Thr Thr Ser Val Asn Met Ser Phe Ala Glu Ala Thr Pro	
20 25 30	
AGT ATT GAA ATG GTT CTT GAT AAA ACT GAA GTC CAT GTA GGA GAT GTA	144
Ser Ile Glu Met Val Leu Asp Lys Thr Glu Val His Val Gly Asp Val	
35 40 45	
ATA ACG GCC ACA ATA AAA GTC AAT AAC ATT AGA AAA TTG GCG GGA TAT	192
Ile Thr Ala Thr Ile Lys Val Asn Asn Ile Arg Lys Leu Ala Gly Tyr	
50 55 60	
CAG CTA AAT ATC AAA TTT GAC CCT GAA GTT TTA CAG CCG GTA GAC CCT	240
Gln Leu Asn Ile Lys Phe Asp Pro Glu Val Leu Gln Pro Val Asp Pro	
65 70 75 80	
GCA ACA GGA GAG GAA TTT ACT GAT AAG TCC ATG CCG GTA AAT AGG GTT	288
Ala Thr Gly Glu Glu Phe Thr Asp Lys Ser Met Pro Val Asn Arg Val	
85 90 95	
TTG CTG ACA AAC AGC AAA TAT GGA CCT ACT CCT GTG GCG GGT AAC GAT	336
Leu Leu Thr Asn Ser Lys Tyr Gly Pro Thr Pro Val Ala Gly Asn Asp	
100 105 110	
ATA AAG TCA GGA ATT ATT AAT TTT GCT ACG GGA TAT AAC AAT TTA ACA	384
Ile Lys Ser Gly Ile Ile Asn Phe Ala Thr Gly Tyr Asn Asn Leu Thr	
115 120 125	



GCG TAC AAA TCC AGC GGA ATA GAC GAA CAT ACA GGA ATA ATA GGA GAG Ala Tyr Lys Ser Ser Gly Ile Asp Glu His Thr Gly Ile Ile Gly Glu 130 135 140	432
ATT GGT TTT AAA GTT TTA AAG AAA CAA AAT ACG TCT ATT AGG TTT GAA Ile Gly Phe Lys Val Leu Lys Lys Gln Asn Thr Ser Ile Arg Phe Glu 145 150 155 160	480
GAT ACA TTA TCG ATG CCC GGG GCA ATA TCG GGA ACA AGT TTG TTT GAC Asp Thr Leu Ser Met Pro Gly Ala Ile Ser Gly Thr Ser Leu Phe Asp 165 170 175	528
TGG GAT GCA GAA ACT ATA ACA GGA TAT GAG GTA ATA CAG CCG GAT CTT Trp Asp Ala Glu Thr Ile Thr Gly Tyr Glu Val Ile Gln Pro Asp Leu 180 185 190	576
ATA GTT GTA GAG GCA GAA CCG TTA AAA GAC GCC AGC GTG GCT CTG GAA Ile Val Val Glu Ala Glu Pro Leu Lys Asp Ala Ser Val Ala Leu Glu 195 200 205	624
CTG GAT AAG ACG AAG GTA AAA GTA GGG GAC ATA ATA ACA GCG ACG ATA Leu Asp Lys Thr Lys Val Lys Val Gly Asp Ile Ile Thr Ala Thr Ile 210 215 220	672
AAG ATA GAG AAC ATG AAG AAT TTT GCA GGG TAC CAG TTG AAT ATC AAG Lys Ile Glu Asn Met Lys Asn Phe Ala Gly Tyr Gln Leu Asn Ile Lys 225 230 235 240	720
TAT GAC CCG ACC ATG TTG GAG GCA ATA GAA CTG GAG ACA GGA AGT GCG Tyr Asp Pro Thr Met Leu Glu Ala Ile Glu Leu Glu Thr Gly Ser Ala 245 250 255	768
ATA GCG AAG AGG ACA TGG CCG GTT ACA GGA GGT ACT GTT CTG CAA AGT Ile Ala Lys Arg Thr Trp Pro Val Thr Gly Gly Thr Val Leu Gln Ser 260 265 270	816
GAC AAT TAT GGA AAG ACG ACT GCG GTA GCG AAT GAT GTA GGA GCA GGT Asp Asn Tyr Gly Lys Thr Thr Ala Val Ala Asn Asp Val Gly Ala Gly 275 280 285	864
ATA ATA AAC TTT GCT GAG GCA TAC TCG AAC CTT ACC AAA TAC AGA GAG Ile Ile Asn Phe Ala Glu Ala Tyr Ser Asn Leu Thr Lys Tyr Arg Glu 290 295 300	912
ACA GGT GTG GCA GAG GAG ACA GGT ATA ATA GGA AAG ATA GGC TTC AGA Thr Gly Val Ala Glu Glu Thr Gly Ile Ile Gly Lys Ile Gly Phe Arg 305 310 315 320	960
GTA CTG AAG GCA GGA AGT ACG GCT ATA AGA TTT GAG GAT ACG ACA GCG Val Leu Lys Ala Gly Ser Thr Ala Ile Arg Phe Glu Asp Thr Thr Ala 325 330 335	1008

ATG CCG GGA GCA ATA GAA GGA ACA TAC ATG TTC GAC TGG TAT GGC GAG Met Pro Gly Ala Ile Glu Gly Thr Tyr Met Phe Asp Trp Tyr Gly Glu 340 345 350	1056
AAC ATC AAA GGG TAT AGC GTA GTA CAG CCT GGG GAA ATA GTG GCA GAA Asn Ile Lys Gly Tyr Ser Val Val Gln Pro Gly Glu Ile Val Ala Glu 355 360 365	1104
GGA GAA GAG CCG GGT GAA GAG CCG ACA GAA GAG CCT GTA CCG ACA GAG Gly Glu Glu Pro Gly Glu Glu Pro Thr Glu Glu Pro Val Pro Thr Glu 370 375 380	1152
ACA CCA GTA GAT CCC ACA CCG ACA GTG ACA GAA GAG CCT GTA CCT TCA Thr Pro Val Asp Pro Thr Pro Thr Val Thr Glu Glu Pro Val Pro Ser 385 390 395 400	1200
GAG CTT CCA GAT TCC TAT GTA ATA ATG GAA CTG GAT AAG ACG AAG GTA Glu Leu Pro Asp Ser Tyr Val Ile Met Glu Leu Asp Lys Thr Lys Val 405 410 415	1248
AAA GTA GGG GAC ATA ATA ACA GCG ACG ATA AAG ATA GAG AAC ATG AAG Lys Val Gly Asp Ile Ile Thr Ala Thr Ile Lys Ile Glu Asn Met Lys 420 425 430	1296
AAT TTT GCA GGG TAC CAG TTG AAT ATC AAG TAT GAC CCG ACC ATG TTG Asn Phe Ala Gly Tyr Gln Leu Asn Ile Lys Tyr Asp Pro Thr Met Leu 435 440 445	1344
GAG GCA ATA GAA CTG GAG ACA GGA AGT GCG ATA GCG AAG AGG ACA TGG Glu Ala Ile Glu Leu Glu Thr Gly Ser Ala Ile Ala Lys Arg Thr Trp 450 455 460	1392
CCG GTT ACA GGA GGT ACT GTT CTG CAA AGT GAC AAT TAT GGA AAG ACG Pro Val Thr Gly Gly Thr Val Leu Gln Ser Asp Asn Tyr Gly Lys Thr 465 470 475 480	1440
ACT GCG GTA GCG AAT GAT GTA GGA GCA GGT ATA ATA AAC TTT GCT GAG Thr Ala Val Ala Asn Asp Val Gly Ala Gly Ile Ile Asn Phe Ala Glu 485 490 495	1488
GCA TAC TCG AAC CTT ACC AAA TAC AGA GAG ACA GGT GTG GCA GAG GAG Ala Tyr Ser Asn Leu Thr Lys Tyr Arg Glu Thr Gly Val Ala Glu Glu 500 505 510	1536
ACA GGT ATA ATA GGA AAG ATA GGC TTC AGA GTA CTG AAG GCA GGA AGT Thr Gly Ile Ile Gly Lys Ile Gly Phe Arg Val Leu Lys Ala Gly Ser 515 520 525	1584
ACG GCT ATA AGA TTT GAG GAT ACG ACA GCG ATG CCG GGA GCA ATA GAA Thr Ala Ile Arg Phe Glu Asp Thr Thr Ala Met Pro Gly Ala Ile Glu 530 535 540	1632

GGA ACA TAC ATG TTC GAC TGG TAT GGC GAG AAC ATC AAA GGG TAT AGC Gly Thr Tyr Met Phe Asp Trp Tyr Gly Glu Asn Ile Lys Gly Tyr Ser 545 550 555 560	1680
GTA GTA CAG CCT GGG GAA ATA GTG GCG GAA GGA GAA GAG CCG ACA GAA Val Val Gln Pro Gly Glu Ile Val Ala Glu Gly Glu Glu Pro Thr Glu 565 570 575	1728
GAG CCT GTA CCG ACA GAG ACA CCA GTA GAT CCC ACA CCG ACA GTG ACA Glu Pro Val Pro Thr Glu Thr Pro Val Asp Pro Thr Pro Thr Val Thr 580 585 590	1776
GAA GAG CCT GTA CCT TCA GAG CTT CCA GAT TCC TAT GTG ATA ATG GAA Glu Glu Pro Val Pro Ser Glu Leu Pro Asp Ser Tyr Val Ile Met Glu 595 600 605	1824
TTG GAT AAG ACG AAG GTA AAA GAA GGC GAC GTA ATA ATA GCA ACA ATA Leu Asp Lys Thr Lys Val Lys Glu Gly Asp Val Ile Ile Ala Thr Ile 610 615 620	1872
AGA GTA AAT AAC ATA AAG AAT CTT GCC GGA TAT CAG ATA GGC ATC AAA Arg Val Asn Asn Ile Lys Asn Leu Ala Gly Tyr Gln Ile Gly Ile Lys 625 630 635 640	1920
TAT GAC CCG AAA GTA TTA GAG GCA TTT AAT ATC GAG ACA GGG GAC CCA Tyr Asp Pro Lys Val Leu Glu Ala Phe Asn Ile Glu Thr Gly Asp Pro 645 650 655	1968
ATA GAT GAA GGA ACA TGG CCT GCA GTA GGG GGA ACA ATA CTG AAG AAT Ile Asp Glu Gly Thr Trp Pro Ala Val Gly Gly Thr Ile Leu Lys Asn 660 665 670	2016
AGA GAT TAC CTG CCG ACT GGG GTA GCA ATA AAC AAT GTA TCT AAA GGA Arg Asp Tyr Leu Pro Thr Gly Val Ala Ile Asn Asn Val Ser Lys Gly 675 680 685	2064
ATA CTG AAT TTT GCT GCT TAT TAC GTT TAC TTC GAT GAC TAT AGA GAG Ile Leu Asn Phe Ala Ala Tyr Tyr Val Tyr Phe Asp Asp Tyr Arg Glu 690 695 700	2112
GAA GGA AAG TCA GAA GAT ACA GGA ATT ATA GGA AAT ATA GGC TTT AGA Glu Gly Lys Ser Glu Asp Thr Gly Ile Ile Gly Asn Ile Gly Phe Arg 705 710 715 720	2160
GTA CTG AAG GCG GAA GAT ACA ACG ATA AGA TTT GAA GAG CTG GAG TCA Val Leu Lys Ala Glu Asp Thr Thr Ile Arg Phe Glu Glu Leu Glu Ser 725 730 735	2208
ATG CCG GGT TCA ATA GAC GGA ACA TAT ATG TTG GAT TGG TAT CTT AAT Met Pro Gly Ser Ile Asp Gly Thr Tyr Met Leu Asp Trp Tyr Leu Asn 740 745 750	2256

AGA ATC TCT GGC TAT GTA GTA ATA CAA CCG GCG CCT ATA AAG GCG GCT Arg Ile Ser Gly Tyr Val Val Ile Gln Pro Ala Pro Ile Lys Ala Ala 755 760 765	2304
AGT GAC GAA CCA ATA CCA ACG GAT ACA CCA TCA GAT GAA CCG ACA CCG Ser Asp Glu Pro Ile Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr Pro 770 775 780	2352
TCA GAC GAG CCA ACG CCA TCT GAC GAA CCG ACA CCG TCT GAT GAG CCA Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro 785 790 795 800	2400
ACA CCG TCA GAT GAA CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile 805 810 815	2448
CCG ACG GAT ACA CCA TCA GAT GAA CCG ACA CCA TCA GAC GAG CCA ACG Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr 820 825 830	2496
CCA TCT GAT GAA CCA ACA CCG TCT GAT GAG CCA ACA CCA TCT GAT GAA Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu 835 840 845	2544
CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro 850 855 860	2592
TCA GAT GAA CCG ACA CCG TCA GAC GAG CCA ACG CCA TCT GAC GAA CCA Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro 865 870 875 880	2640
ACA CCG TCT GAT GAG CCA ACA CCG TCA GAT GAA CCG ACT CCG TCA GAG Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu 885 890 895	2688
ACA CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA TCA GAT GAA CCG ACA Thr Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr 900 905 910	2736
CCG TCA GAC GAG CCA ACG CCA TCT GAC GAA CCA ACA CCG TCT GAT GAG Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu 915 920 925	2784
CCA ACA CCG TCA GAT GAA CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro 930 935 940	2832
ATA CCG ACG GAT ACA CCA TCA GAT GAA CCG ACA CCG TCA GAC GAG CCG Ile Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro 945 950 955 960	2880

ACA CCA TCT GAC GAA CCA ACA CCG TCA GAC GAG CCA ACG CCA TCT GAC Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp 965 970 975	2928
GAA CCG ACA CCG TCT GAT GAG CCA ACA CCA TCT GAT GAA CCG ACT CCG Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro 980 985 990	2976
TCA GAG ACA CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA TCA GAT GAA Ser Glu Thr Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro Ser Asp Glu 995 1000 1005	3024
CCG ACA CCG TCA GAC GAG CCG ACA CCA TCT GAC GAA CCA ACA CCG TCA Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser 1010 1015 1020	3072
GAC GAG CCA ACG CCA TCT GAC GAA CCG ACA CCG TCT GAT GAG CCA ACA Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr 1025 1030 1035 1040	3120
CCA TCT GAT GAA CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA CCG Pro Ser Asp Glu Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile Pro 1045 1050 1055	3168
ACG GAT ACA CCA TCA GAT GAA CCG ACA CCG TCA GAC GAG CCG ACA CCA Thr Asp Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro 1060 1065 1070	3216
TCT GAC GAA CCA ACA CCG TCT GAT GAG CCA ACA CCG TCA GAT GAA CCG Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro 1075 1080 1085	3264
ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA TCA Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro Ser 1090 1095 1100	3312
GAT GAA CCG ACA CCG TCA GAC GAG CCA ACG CCA TCT GAC GAA CCG ACA Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr 1105 1110 1115 1120	3360
CCG TCT GAT GAG CCA ACA CCG TCA GAT GAA CCG ACT CCG TCA GAG ACA Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu Thr 1125 1130 1135	3408
CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA TCA GAT GAA CCG ACA CCG Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr Pro 1140 1145 1150	3456
TCA GAC GAG CCA ACG CCA TCT GAC GAA CCG ACA CCG TCT GAT GAG CCA Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro 1155 1160 1165	3504

ACA CCG TCA GAT GAA CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile 1170 1175 1180	3552
CCG ACG GAT ACA CCA TCA GAT GAA CCG ACA CCA TCA GAC GAG CCA ACG Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr 1185 1190 1195 1200	3600
CCA TCT GAT GAA CCA ACA CCG TCT GAT GAG CCA ACA CCA TCT GAT GAA Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu 1205 1210 1215	3648
CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro 1220 1225 1230	3696
TCA GAT GAA CCG ACA CCG TCA GAC GAG CCA ACG CCA TCT GAC GAA CCA Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro 1235 1240 1245	3744
ACA CCG TCT GAT GAG CCA ACA CCG TCA GAT GAA CCG ACT CCG TCA GAG Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu 1250 1255 1260	3792
ACA CCT GAG GAG CCG ATA CCG ACG GAT ACA CCA TCA GAT GAA CCG ACA Thr Pro Glu Glu Pro Ile Pro Thr Asp Thr Pro Ser Asp Glu Pro Thr 1265 1270 1275 1280	3840
CCG TCA GAC GAG CCG ACA CCA TCT GAC GAA CCA ACA CCG TCA GAC GAG Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu 1285 1290 1295	3888
CCA ACG CCA TCT GAC GAA CCG ACA CCG TCT GAT GAG CCA ACA CCA TCT Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser 1300 1305 1310	3936
GAT GAA CCG ACT CCG TCA GAG ACA CCT GAG GAG CCG ATA CCG ACG GAT Asp Glu Pro Thr Pro Ser Glu Thr Pro Glu Glu Pro Ile Pro Thr Asp 1315 1320 1325	3984
ACA CCA TCA GAT GAA CCG ACA CCG TCA GAC GAG CCG ACA CCA TCT GAC Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp 1330 1335 1340	4032
GAA CCA ACA CCG TCA GAC GAG CCA ACG CCA TCT GAC GAA CCG ACA CCG Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro 1345 1350 1355 1360	4080
TCT GAT GAG CCA ACA CCA TCT GAT GAA CCG ACT CCG TCA GAG ACA CCT Ser Asp Glu Pro Thr Pro Ser Asp Glu Pro Thr Pro Ser Glu Thr Pro 1365 1370 1375	4128

GAG GAG CCG ACA CCG ACT ACT ACA CCG ACA CCA ACA CCG TCG ACA ACG Glu Glu Pro Thr Pro Thr Thr Thr Pro Thr Pro Thr Pro Ser Thr Thr 1380 1385 1390	4176
CCT ACA AGT GGC AGC GGA GGC AGT GGT GGA AGC GGT GGT GGC GGC GGA Pro Thr Ser Gly Ser Gly Gly Ser Gly Gly Ser Gly Gly Gly Gly Gly 1395 1400 1405	4224
GGT GGT GGA GGA ACT GTA CCT ACA TCT CCA ACA CCG ACA CCG ACA TCT Gly Gly Gly Gly Thr Val Pro Thr Ser Pro Thr Pro Thr Pro Thr Ser 1410 1415 1420	4272
AAA CCG ACG TCT ACA CCT GCA CCG ACA GAA ATC GAA GAG CCT ACA CCA Lys Pro Thr Ser Thr Pro Ala Pro Thr Glu Ile Glu Glu Pro Thr Pro 1425 1430 1435 1440	4320
TCT GAT GTG CCT GGT GCA ATC GGT GGA GAA CAT AGA GCA TAC TTA AGA Ser Asp Val Pro Gly Ala Ile Gly Gly Glu His Arg Ala Tyr Leu Arg 1445 1450 1455	4368
GGA TAT CCG GAT GGA AGC TTC AGG CCT GAA AGA AAT ATA ACA AGA GCT Gly Tyr Pro Asp Gly Ser Phe Arg Pro Glu Arg Asn Ile Thr Arg Ala 1460 1465 1470	4416
GAA GCG GCG GTA ATC TTT GCT AAG TTG CTT GGA GCC GAT GAA AGC TAT Glu Ala Ala Val Ile Phe Ala Lys Leu Leu Gly Ala Asp Glu Ser Tyr 1475 1480 1485	4464
GGA GCT CAG TCT GCA AGT CCA TAT AGT GAT TTG GCT GAT ACT CAC TGG Gly Ala Gln Ser Ala Ser Pro Tyr Ser Asp Leu Ala Asp Thr His Trp 1490 1495 1500	4512
GCT GCA TGG GCA ATC AAA TTT GCA ACA AGC CAG GGC TTG TTC AAA GGA Ala Ala Trp Ala Ile Lys Phe Ala Thr Ser Gln Gly Leu Phe Lys Gly 1505 1510 1515 1520	4560
TAT CCG GAC GGT ACG TTT AAA CCT GAT CAG AAC ATA ACG AGA GCG GAA Tyr Pro Asp Gly Thr Phe Lys Pro Asp Gln Asn Ile Thr Arg Ala Glu 1525 1530 1535	4608
TTC GCA ACT GTG GTA CTC CAC TTC CTG ACA AAA GTT AAG GGT CAG GAA Phe Ala Thr Val Val Leu His Phe Leu Thr Lys Val Lys Gly Gln Glu 1540 1545 1550	4656
ATA ATG AGC AAG CTT GCA ACA ATA GAT ATA AGT AAT CCG AAG TTT GAC Ile Met Ser Lys Leu Ala Thr Ile Asp Ile Ser Asn Pro Lys Phe Asp 1555 1560 1565	4704
GAT TGT GTC GGA CAT TGG GCA CAA GAG TTT ATT GAG AAA TTG ACA AGC Asp Cys Val Gly His Trp Ala Gln Glu Phe Ile Glu Lys Leu Thr Ser 1570 1575 1580	4752





GAC ATA ATT ATA GCC ACA ATA AGA ATT GAC AAT ATC AAT AAC TTT AGC Asp Ile Ile Ile Ala Thr Ile Arg Ile Asp Asn Ile Asn Asn Phe Ser 50 55 60	192
GGA TAT CAA TTA AAT ATA AAG TAT GAT CCG TCA TAC CTC CAG GCA GTT Gly Tyr Gln Leu Asn Ile Lys Tyr Asp Pro Ser Tyr Leu Gln Ala Val 65 70 75 80	240
AAT CCT TTG ACA GGA GAA CCG ATA AAA AAG AGA ACA ATG CCG GCA GTG Asn Pro Leu Thr Gly Glu Pro Ile Lys Lys Arg Thr Met Pro Ala Val 85 90 95	288
AAC GGC ACG GTG TTG TTA AAG GGA GAT CAG TAC AGT ATT ACT GAG GTT Asn Gly Thr Val Leu Leu Lys Gly Asp Gln Tyr Ser Ile Thr Glu Val 100 105 110	336
GTA GAA AAT AAC GTC GAT GAA GGG ATT TTA AAT TTT GGC AAG GGA TAT Val Glu Asn Asn Val Asp Glu Gly Ile Leu Asn Phe Gly Lys Gly Tyr 115 120 125	384
GCA AAT TTA ACT GAA TAC AGG AAA AGC GGA AAA CCT GAA ACA ACC GGA Ala Asn Leu Thr Glu Tyr Arg Lys Ser Gly Lys Pro Glu Thr Thr Gly 130 135 140	432
ATT ATT GGC AAG ATA GGA TTT AAA GCC TTA AAG CTT GGC AAG ACG GAG Ile Ile Gly Lys Ile Gly Phe Lys Ala Leu Lys Leu Gly Lys Thr Glu 145 150 155 160	480
ATC AAA TTT GAG AAC ACA CCC GTC ATG CCT GGG GCA AAA GAA GGA ACA Ile Lys Phe Glu Asn Thr Pro Val Met Pro Gly Ala Lys Glu Gly Thr 165 170 175	528
CTG CTG TTT GAC TGG GAT GCA GAA ACT ATA ACG GAA TAT AAT GTA ATT Leu Leu Phe Asp Trp Asp Ala Glu Thr Ile Thr Glu Tyr Asn Val Ile 180 185 190	576
CAG CCT AAA GAA CTT GCA ATA ACG TTA CCG GAC GAT GCA CAC ATT GCT Gln Pro Lys Glu Leu Ala Ile Thr Leu Pro Asp Asp Ala His Ile Ala 195 200 205	624
TTG GAA CTT GAC AAG ACA AAA GTG AAA GTG GGA GAT GTA ATT GTT GCG Leu Glu Leu Asp Lys Thr Lys Val Lys Val Gly Asp Val Ile Val Ala 210 215 220	672
ACA GTA AAA GCA AAG AAT ATG ACT AGT ATG GCG GGA ATT CAG GTA AAT Thr Val Lys Ala Lys Asn Met Thr Ser Met Ala Gly Ile Gln Val Asn 225 230 235 240	720
ATT AAA TAT GAC CCT GAA GTA TTG CAG GCG ATT GAT CCT GCG ACG GGA Ile Lys Tyr Asp Pro Glu Val Leu Gln Ala Ile Asp Pro Ala Thr Gly 245 250 255	768

AAA CCG TTT ACA AAA GAA ACA TTA CTT GTG GAC CCG GAA CTG TTA TCA Lys Pro Phe Thr Lys Glu Thr Leu Leu Val Asp Pro Glu Leu Leu Ser 260 265 270	816
AAC AGA GAA TAT AAT CCG TTG TTA ACA GCA GTT AAT GAC ATA AAT TCC Asn Arg Glu Tyr Asn Pro Leu Leu Thr Ala Val Asn Asp Ile Asn Ser 275 280 285	864
GGC ATT ATA AAT TAT GCA TCT TGT TAT GTA TAT TGG GAT TCC TAC AGA Gly Ile Ile Asn Tyr Ala Ser Cys Tyr Val Tyr Trp Asp Ser Tyr Arg 290 295 300	912
GAA TCA GGA GTA TCT GAA AGC ACC GGA ATA ATT GGA AAG GTT GGC TTT Glu Ser Gly Val Ser Glu Ser Thr Gly Ile Ile Gly Lys Val Gly Phe 305 310 315 320	960
AAA GTG CTG AAA GCT GCC AAC ACC ACA GTA AAA CTG GAA GAA ACA AGA Lys Val Leu Lys Ala Ala Asn Thr Thr Val Lys Leu Glu Glu Thr Arg 325 330 335	1008
TTT ACA CCA AAT TCG ATA GAC GGT ACT TTG GTA ATT GAT TGG TAT GGC Phe Thr Pro Asn Ser Ile Asp Gly Thr Leu Val Ile Asp Trp Tyr Gly 340 345 350	1056
CAA CAG ATA GTT GGT TAT AAA GTA ATA CAG CCC GAC AAA ATT ACT GTG Gln Gln Ile Val Gly Tyr Lys Val Ile Gln Pro Asp Lys Ile Thr Val 355 360 365	1104
ATT TCA GAG CCT GAG GTA CCA ACA CAA ACA CCT ACA CAG ACA CCG CCA Ile Ser Glu Pro Glu Val Pro Thr Gln Thr Pro Thr Gln Thr Pro Pro 370 375 380	1152
ACA ACA ACA GCA CCA TCG CAA ACA CCT ACG CAG ACA CCG CCA ACA ACA Thr Thr Thr Ala Pro Ser Gln Thr Pro Thr Gln Thr Pro Pro Thr Thr 385 390 395 400	1200
ACA GCA CCA TCA CAG ACA CCT ACA CAG ACA CCG GCA GTA ACG CCG ACG Thr Ala Pro Ser Gln Thr Pro Thr Gln Thr Pro Ala Val Thr Pro Thr 405 410 415	1248
CAA AGT GCA ACT CCG TCG GAT CCT GGC GGA GGT GGA GGA GGC CTC CCG Gln Ser Ala Thr Pro Ser Asp Pro Gly Gly Gly Gly Gly Gly Leu Pro 420 425 430	1296
GGT GGT GGA GGC GGC GCT GTT AAT CCT TCA GCT TCA CCG ACA CCA ACA Gly Gly Gly Gly Ala Val Asn Pro Ser Ala Ser Pro Thr Pro Thr 435 440 445	1344
CCG ACA TCC AAA CCT ACT CCT ACT GCC ACT AAA AAA CCG GAG CCA ACG Pro Thr Ser Lys Pro Thr Pro Thr Ala Thr Lys Lys Pro Glu Pro Thr 450 455 460	1392

GAA ATA GAA GAA CCC GAA CCT GAA ATA CCG GGC ACT GTT GGA ATA CAT Glu Ile Glu Glu Pro Glu Pro Glu Ile Pro Gly Thr Val Gly Ile His 465 470 475 480	1440
TAT TCA TAC CTG ACA GGT TAT CCG GAC AAA ATG TTC AGA CCT GAA AAG Tyr Ser Tyr Leu Thr Gly Tyr Pro Asp Lys Met Phe Arg Pro Glu Lys 485 490 495	1488
AGT ATT ACA AGA GCT GAA GCA GCC GTG ATT TTT GCA AAA CTT TTG GGA Ser Ile Thr Arg Ala Glu Ala Ala Val Ile Phe Ala Lys Leu Leu Gly 500 505 510	1536
GCA AAC GAA AAT ACA AAG ATA AAC TAT AAT GTT TCA TAC ACC GAT GTT Ala Asn Glu Asn Thr Lys Ile Asn Tyr Asn Val Ser Tyr Thr Asp Val 515 520 525	1584
GAC AGC TCC CAT TGG GCA AGT TGG GCA ATC AAA TTT GTA TCA TAC AAG Asp Ser Ser His Trp Ala Ser Trp Ala Ile Lys Phe Val Ser Tyr Lys 530 535 540	1632
AAA CTG TTT ACC GGA TAT CCT GAT GGC TCG TTC AAG CCT AAT CAG AAT Lys Leu Phe Thr Gly Tyr Pro Asp Gly Ser Phe Lys Pro Asn Gln Asn 545 550 555 560	1680
ATA ACG AGA GCC GAA TTT TCA ACG GTT GTG TTT AAG CTT CTT GTA TCT Ile Thr Arg Ala Glu Phe Ser Thr Val Val Phe Lys Leu Leu Val Ser 565 570 575	1728
GAG AAA GGT CTA AAA GAA GAA AAG ATT GAA AAG TCC AAG TTT GGT GAT Glu Lys Gly Leu Lys Glu Glu Lys Ile Glu Lys Ser Lys Phe Gly Asp 580 585 590	1776
ACA AAG GGC CAC TGG GCA CAA CAG TTT ATT GAA CAG CTG TCA GAC CTT Thr Lys Gly His Trp Ala Gln Gln Phe Ile Glu Gln Leu Ser Asp Leu 595 600 605	1824
GGA TAC ATC AAC GGA TAT CCT GAT GGT ACA TTC AAG CCC AAC AAC AAT Gly Tyr Ile Asn Gly Tyr Pro Asp Gly Thr Phe Lys Pro Asn Asn Asn 610 615 620	1872
ATC AAA CGA TCA GAA AGT GTT GCC CTG ATA AAC AGA GCT ATG GGA AGA Ile Lys Arg Ser Glu Ser Val Ala Leu Ile Asn Arg Ala Met Gly Arg 625 630 635 640	1920
GGG CCT TTG CAT GGC GCA CCG CAG GTA TTC GAG GAT GTT CCT CAG ACA Gly Pro Leu His Gly Ala Pro Gln Val Phe Glu Asp Val Pro Gln Thr 645 650 655	1968
CAC TGG GCT TTC AAA GAT ATT GCA GAG GGC GTG CTC AAT CAC AGA TAC His Trp Ala Phe Lys Asp Ile Ala Glu Gly Val Leu Asn His Arg Tyr 660 665 670	2016

AAA CTG GAC AAT GAG GGC AAA GAA CAA TTG CTG GAG ATA ATT GAT AAC 2064

Lys Leu Asp Asn Glu Gly Lys Glu Gln Leu Leu Glu Ile Ile Asp Asn  
675 680 685

DESCRIPTION DE LA SEQUENCE : SEQ ID N° : 4:

**SEQUENCE NUCLEOTIDIQUE DE LA PROTEINE CipA**

ATGAGAAAAGTCATCAGTATGCTCTTAGTTGTGGCTATGCTGACGACGATTTTTGCGGCGATGATAC  
CGC  
AGACAGTATCGGCGGCCACAATGACAGTCGAGATCGGCAAAGTTACAGCAGCCGTTGGATCAAAA  
GTAGA  
AATACCTATAACCCTGAAAGGAGTGCCATCCAAAGGAATGGCCAATTGCGACTTCGTATTGGGT  
TGAT  
CCAAATGTGCTGGAAGTAACAGAAGTAAAACCAGGAAGCATAATAAAAGATCCGGATCCTAGCAA  
GAGCT  
TTGATAGCGCAATATATCCGGATCGAAAGATGATTGTATTTCTGTTTGCAGAAGACAGTGGAAGAG  
GAAC  
GTATGCAATAAATCAGGATGGAGTATTTGCAACAATTGTAGCCACTGTCAAATCAGCTGCAGCGGC  
ACCG  
ATTACTTTGCTTGAAGTAGGTGCATTTGCGGACAACGATTTAGTAGAAATAAGCACAACTTTTGTCC  
CGG  
GCGGAGTAAATCTTGGTAGTTCCGTACCGACAACACAGCCAAATGTTCCGTCAGACGGTGTGGTAG  
TAGA  
AATTGGCAAAGTTACGGGATCTGTTGGAACACAGTTGAAATACCTGTATATTTAGAGGAGTTCC  
ATCC  
AAAGGAATAGCAAACCTGCGACTTTGTGTTTCAAGATATGATCCGAATGTATTGGAAATTATAGGGATA  
GATC  
CCGGAGACATAATAGTTGACCCGAATCCTACCAAGAGCTTTGATACTGCAATATATCCTGACAGAA  
AGAT  
AATAGTATTCCTGTTTGCAGGAAGACAGCGGAACAGGAGCGTATGCAATAACTAAAGACGGAGTATT  
TGCA  
AAAATAAGAGCAACTGTAAATCAAGTGCTCCGGGCTATATTACTTTGACGAAGTAGGTGGATTT  
GCAG  
ATAATGACCTGGTAGAACAGAAGGTATCATTTATAGACGGTGGTGTAAACGTTGGCAATGCAACAC  
CGAC  
CAAGGGAGCAACACCAACAAATACAGCTACGCCGACAAAATCAGCTACGGCTACGCCACACAGGC  
CATCG  
GTACCGACAAACACACCGACAAACACACCGGCAAATACACCGGTATCAGGCAATTTGAAGGTTGA  
ATTCT  
ACAACAGCAATCCTTCAGATACTACTAATCAATCAATCCTCAGTTCAAGGTTACTAATACCGGAA  
GCAG  
TGCAATTGATTTGTCCAACTCACATTGAGATATTATTATACAGTAGACGGACAGAAAGATCAGAC  
CTTC  
TGGTGTGACCATGCTGCAATAATCGGCAGTAACGGCAGCTACAACGGAATTACTTCAAATGTAAAA  
GGAA  
CATTTGTAAAAATGAGTTCCTCAACAAATAACGCAGACACCTACCTTGAAATAAGCTTTACAGGCG  
GAAC  
TCTTGAACCGGGTGACATGTTTCAAGGTAGATTGCAAAGAATGACTGGAGTAACTATAC  
ACAG  
TCAAATGACTACTCATTCAAGTCTGCTTCACAGTTTGTGTAATGGGATCAGGTAACAGCATACTGA  
ACG  
GTGTTCTTGATGGGGTAAAGAACCCGGTGGCAGTGATACCATCAACACAGCCTGTAACAACAC  
CACC  
TGCAACAACAAAACCACTGCAACAACAAAACCACTGCAACAACAATACCGCCGTCAGATGATCC  
GAAT  
GCAATAAAGATTAAGGTGGACACAGTAAATGCAAAACCGGGAGACACAGTAAATATACCTGTAAG  
ATTCA  
GTGGTATACCATCCAAGGGAATAGCAAACCTGTGACTTTGTATACAGCTATGACCCGAATGTA  
AGAT

AATAGAGATAAAACCGGGAGAATTGATAGTTGACCCGAATCCTGACAAGAGCTTTGATACTGCAGT  
ATAT  
CCTGACAGAAAGATAATAGTATTCTGTTTGCAGAAGACAGCGGAACAGGAGCGTATGCAATAACT  
AAAG  
ACGGAGTATTTGCTACGATAGTAGCGAAAGTAAAATCCGGAGCACCTAACGGACTCAGTGTAATCA  
AATT  
TGTAAGAAGTAGGCGGATTTGCGAACAATGACCTTGTAAGACAGAGGACACAGTTCTTTGACGGTGG  
AGTA  
AATGTTGGAGATACAACAGTACCTACAACACCTACAACACCTGTAACAACACCGACAGATGATTGG  
AATG  
CAGTAAGGATTAAGGTGGACACAGTAAATGCAAAACCGGGAGACACAGTAAGAATACCTGTAAGA  
TTCAG  
CGGTATACCATCCAAGGGAATAGCAAACCTGTGACTTTGTATACAGCTATGACCCGAATGTACTTGA  
GATA  
ATAGAGATAGAACCGGGAGACATAATAGTTGACCCGAATCCTGACAAGAGCTTTGATACTGCAGTA  
TATC  
CTGACAGAAAGATAATAGTATTCTGTTTGCAGAAGACAGCGGAACAGGAGCGTATGCAATAACTA  
AAGA  
CGGAGTATTTGCTACGATAGTAGCGAAAGTAAAATCCGGAGCACCTAACGGACTCAGTGTAATCAA  
ATTT  
GTAGAAGTAGGCGGATTTGCGAACAATGACCTTGTAAGACAGAAGACACAGTTCTTTGACGGTGG  
GTAA  
ATGTTGGAGATACAACAGAACCTGCAACACCTACAACACCTGTAACAACACCGACAACAACAGAT  
GATCT  
GGATGCAGTAAGGATTAAGGTGGACACAGTAAATGCAAAACCGGGAGACACAGTAAGAATACCTG  
TAAGA  
TTCAGCGGTATACCATCCAAGGGAATAGCAAACCTGTGACTTTGTATACAGCTATGACCCGAATGTA  
CTTG  
AGATAATAGAGATAGAACCGGGAGACATAATAGTTGACCCGAATCCTGACAAGAGCTTTGATACTG  
CAGT  
ATATCCTGACAGAAAGATAATAGTATTCTGTTTGCAGAAGACAGCGGAACAGGAGCGTATGCAAT  
AACT  
AAAGACGGAGTATTTGCTACGATAGTAGCGAAAGTAAAATCCGGAGCACCTAACGGACTCAGTGT  
AATCA  
AATTTGTAGAAGTAGGCGGATTTGCGAACAATGACCTTGTAAGACAGAAGACACAGTTCTTTGACG  
GTGG  
AGTAAATGTTGGAGATACAACAGAACCTGCAACACCTACAACACCTGTAACAACACCGACAACAA  
CAGAT  
GATCTGGATGCAGTAAGGATTAAGGTGGACACAGTAAATGCAAAACCGGGAGACACAGTAAGAAT  
ACCTG  
TAAGATTCAGCGGTATACCATCCAAGGGAATAGCAAACCTGTGACTTTGTATACAGCTATGACCCGA  
ATGT  
ACTTGAGATAATAGAGATAGAACCGGGAGACATAATAGTTGACCCGAATCCTGACAAGAGCTTTGA  
TACT  
GCAGTATATCCTGACAGAAAGATAATAGTATTCTGTTTGCAGAAGACAGCGGAACAGGAGCGTAT  
GCAA  
TAACTAAAGACGGAGTATTTGCTACGATAGTAGCGAAAGTAAAAGAAGGAGCACCTAACGGACTC  
AGTGT  
AATCAAATTTGTAGAAGTAGGCGGATTTGCGAACAATGACCTTGTAAGACAGAAGACACAGTTCTT  
TGAC  
GGTGGAGTAAATGTTGGAGATACAACAGAACCTGCAACACCTACAACACCTGTAACAACACCGAC  
AACA  
CAGATGATCTGGATGCAGTAAGGATTAAGGTGGACACAGTAAATGCAAAACCGGGAGACACAGTA  
AGAAT  
ACCTGTAAGATTCAGCGGTATACCATCCAAGGGAATAGCAAACCTGTGACTTTGTATACAGCTATGA  
CCCG  
AATGTACTTGAGATAATAGAGATAGAACCGGGAGAATTGATAGTTGACCCGAATCCTACCAAGAGC  
TTTG

ATACTGCAGTATATCCTGACAGAAAGATGATAGTATTCCTGTTTGC GGAAGACAGCGGAACAGGAG  
CGTA  
TGCAATAACTGAAGATGGAGTATTTGCTACGATAGTAGCGAAAGTAAATCCGGAGCACCTAACGG  
ACTC  
AGTGTAATCAAATTTGTAGAAGTAGGCGGATTTGCGAACAATGACCTTGTAAGACAGAAGACACAG  
TTCT  
TTGACGGTGGAGTAAATGTTGGAGATACAACAGAACCTGCAACACCTACAACACCTGTAACAACAC  
CGAC  
AACAACAGATGATCTGGATGCAGTAAGGATTAAAGTGGACACAGTAAATGCAAAACCGGGAGACA  
CAGTA  
AGAATACCTGTAAGATTCAGCGGTATACCATCCAAGGGAATAGCAAACCTGTGACTTTGTATACAGC  
TATG  
ACCCGAATGTACTTGAGATAATAGAGATAGAACCGGGAGACATAATAGTTGACCCGAATCCTGACA  
AGAG  
CTTGATACTGCAGTATATCCTGACAGAAAGATAATAGTATTCCTGTTTGCAGAAGACAGCGGAAC  
GGGA  
GCGTATGCAATAACTAAAGACGGAGTATTTGCTACGATAGTAGCGAAAGTAAAAGAAGGAGCACC  
TAACG  
GACTCAGTGTAATCAAATTTGTAGAAGTAGGCGGATTTGCGAACAATGACCTTGTAAGACAGAAGA  
CACA  
GTTCTTTGACGGTGGAGTAAATGTTGGAGATACAACAGTACCTACAACATCGCCGACAACAACACC  
GCCA  
GAGCCGACGATAACTCCGAACAAGTTGACACTTAAGATAGGCAGAGCAGAAGGAAGACCTGGAGA  
CACGG  
TGGAATACCGGTTAACTTGTATGGAGTACCTCAAAAAGGAATAGCAAGCGGTGACTTCGTAGTAA  
GCTA  
TGACCCGAATGTACTTGAGATAATAGAGATAGAACCGGGAGAATTGATAGTTGACCCGAATCCTAC  
CAAG  
AGCTTTGATACTGCAGTATATCCTGACAGAAAGATGATAGTATTCCTGTTTGC GGAAGACAGCGGA  
ACAG  
GAGCGTATGCAATAACTGAAGATGGAGTATTTGCTACGATAGTAGCGAAAGTAAAAGAAGGAGCA  
CCTGA  
AGGATTCAGTGCAATAGAAATTTCTGAGTTTGGTGCATTTGCAGATAATGATCTGGTAGAAGTGGA  
AACT  
GACCTTATCAATGGTGGAGTACTTGTAACATAAAACCTGTAATAGAAGGATATAAAGTATCCGGA  
TACA  
TTTTGCCAGACTTCTCCTTCGACGCTACTGTTGCACCACTTGTAAGGCCGGATTCAAAGTTGAAAT  
AGT  
AGGAACAGAATTGTATGCAGTAACAGATGCAAACGGATACTTTGAAATAACCGGAGTACCTGCAA  
ATGCA  
AGCGGATATACATTGAAGATTTCAAGAGCAACTTACTTGGACAGAGTAATTGCAAATGTTGTAGTA  
ACGG  
GAGATACTTCAGTTTCAACTTCACAGGCTCCAATAATGATGTGGGTAGGAGACATAGTGAAAGACA  
ATTG  
TATCAACCTGTTGGACGTTGCAGAAGTTATCCGTTGCTTCAACGCTACTAAAGGAAGCGCAAACTA  
CGTA  
GAAGAACTTGACATTAATAGAAACGGCGCAATTAACATGCAAGACATAATGATTGTTTCATAAGCAC  
TTTG  
GAGCTACATCAAGTGATTACGACGCACAGTAA

**SEQUENCE DE LA PROTEINE CipA**

MRKVISMMLLVVAMLTTFIFAAMIPQTVSAATMTVEIGKVTAAVGSKVEIPITLKGVP SKGMANCDFVLGY  
DPNVLEVTEVKPGSIIKDPDPSKSFDSAIPDRKMIVFLFAEDSGRGTYAITQDGVFATIVATVKSAAAAP  
ITLLEVGAFAADNDLVEISTTFVAGGVNLGSSVPTTQPNVPSDGVVVEIGKVTGSGVTTVEIPVYFRGVPS  
KGIANCDFVFRYDPNVLEIIGIDPGDIIVDPNPTKSFDTAIYPDRKIIVFLFAEDSGTGAYAITKDGVF AKIR  
ATVKSSAPGYITFDEVGGFADNDLVEQKVSFIDGGVNVGNATPTKGATPTNTATPTKSATATPTRPSVPT  
NTPNTNPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNTGSSAIDLSKLT LRYYYTVDGQKDQTFWC  
DHAAIIGSNGSYNGITSNVKGT FVKMSSSTNNADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQS  
NDYSFKSASQFVEWDQVTAYLNGVLVWGKEPGGSVVPSTQPVTTTPATT KPPATT KPPATTIPPSDDPN  
AIKIKVDTVNAKPGDTVNI PVRFSGIPSKGIANCDFVYSYDPNVLEIIEIKPGELIVDPNPKSFDTAVYPD  
RKIIVFLFAEDSGTGAYAITKDGVFATIVAKVKSGAPNGLSVIKFVEVGGFANNDLVEQRTQFFDGGVN  
VGDTTVPTTPTTPTTPTTPTTDDSN AVRIKVDTVNAKPGDTVRI PVRFSGIPSKGIANCDFVYSYDPNVLEIIEI  
EPGDIIVDPNPKSFDTAVYPDRKIIVFLFAEDSGTGAYAITKDGVFATIVAKVKSGAPNGLSVIKFVEVG  
GFANNDLVEQKTQFFDGGVNVGDTTEPATPTTPTTPTTDDLD AVRIKVDTVNAKPGDTVRI PVRFSG  
IPSKGIANCDFVYSYDPNVLEIIEIEPGDIIVDPNPKSFDTAVYPDRKIIVFLFAEDSGTGAYAITKDGVFA  
TIVAKVKSGAPNGLSVIKFVEVGGFANNDLVEQKTQFFDGGVNVGDTTEPATPTTPTTPTTDDLD AV  
RIKVDTVNAKPGDTVRI PVRFSGIPSKGIANCDFVYSYDPNVLEIIEIEPGDIIVDPNPKSFDTAVYPDRKI  
IVFLFAEDSGTGAYAITKDGVFATIVAKVKEGAPNGLSVIKFVEVGGFANNDLVEQKTQFFDGGVNVGD  
TTEPATPTTPTTPTTPTTDDLD AVRIKVDTVNAKPGDTVRI PVRFSGIPSKGIANCDFVYSYDPNVLEIIEIEP  
GELIVDPNPTKSFDTA VYPDRKMIVFLFAEDSGTGAYAITEDGVFATIVAKVKSGAPNGLSVIKFVEVGG  
FANNDLVEQKTQFFDGGVNVGDTTEPATPTTPTTPTTDDLD AVRIKVDTVNAKPGDTVRI PVRFSGIP  
SKGIANCDFVYSYDPNVLEIIEIEPGDIIVDPNPKSFDTAVYPDRKIIVFLFAEDSGTGAYAITKDGVFATI  
VAKVKEGAPNGLSVIKFVEVGGFANNDLVEQKTQFFDGGVNVGDTTVPTTSP TTTTPEPTITPNKLT LKI  
GRAEGRPGDTVEIPVNLYGVPQKGIASGDFVYSYDPNVLEIIEIEPGELIVDPNPTKSFDTA VYPDRKMIV  
FLFAEDSGTGAYAITEDGVFATIVAKVKEGAPEGFSAIESEFGAFADNDLVEVETDLINGGVLVTNKPVI  
EGYKVSGYILPDFSF DATVAPLVKAGFKVEIVGTELYAVTDANGYFEITGVPANASGYTLKISRATY LDR  
VIANVVVTGDTSVSTSQAPIMMWVGDIVKDNSINLLDVAEVIRCFNATKGSANYVEELDINRNGAINMQ  
DIMIVHKHFGATSSDYDAQ



REVENDICATIONS

1. Composé sur lequel est capable de se fixer de façon covalente ou non au moins un domaine cohésine de type II.
- 5 2. Composé selon la revendication 1, caractérisé en ce que le domaine cohésine de type II provient d'une bactérie cellulolytique.
3. Composé selon la revendication 2, caractérisé en ce que le domaine cohésine de type II provient d'une souche de Clostridium, et notamment de Clostridium thermocellum.
- 10 4. Composé selon la revendication 3, caractérisé en ce que le domaine cohésine de type II provient d'une protéine de Clostridium thermocellum ou d'un fragment de celle-ci compris entre 50 et 600 acides aminés.
- 15 5. Composé selon la revendication 3, caractérisé en ce que le domaine cohésine de type II provient d'une protéine de Clostridium thermocellum choisie parmi SdbA, OlpB et ORF2p ou d'une protéine homologue.
- 20 6. Composé selon la revendication 5, caractérisé en ce qu'il comprend la séquence de 165 acides aminés substantiellement telle que représentée dans l'IDS n° 1 de l'acide aminé n° 27 à l'acide aminé n° 210 de la séquence de la protéine SdbA ou une séquence homologue ou un fragment de cette séquence ou d'une séquence homologue ayant une activité cohésine de type II.
- 25 7. Composé selon la revendication 5, caractérisé en ce qu'il comprend comme domaine cohésine de type II l'une des séquences de la protéine OlpB choisies parmi la séquence des acides aminés n° 28 au n° 190, la séquence des acides aminés n° 207 au n° 362, la séquence des acides aminés n° 409 au n° 564 et la séquence des acides aminés n° 607 au n° 762 de l'IDS n° 2 ou une séquence homologue à l'une de ces séquences ou un
- 30 fragment de ces séquences d'au moins 50 acides aminés ayant une activité cohésine de type II.
- 35 8. Composé selon la revendication 5, caractérisé en ce qu'il a un domaine cohésine de type II, une séquence de la protéine ORF2p choisie parmi la séquence des acides aminés n° 38 à 194 et la séquence des acides aminés n° 209 à 364 de l'IDS n° 3, ou une séquence homologue à ces séquences ou un fragment de ces séquences d'au moins 50 acides aminés ayant une activité cohésine de type II

9. Composé selon l'une des revendications 1 à 8, caractérisé en ce qu'il s'agit essentiellement d'un polypeptide ou d'une protéine.

10. Composé selon la revendication 9, caractérisé en ce qu'il s'agit d'une protéine à activité enzymatique.

5 11. Composé selon l'une des revendications 1 à 10, caractérisé en ce qu'il comporte au moins un autre domaine cohésine qui n'est pas de type II et/ou un domaine dockerine.

10 12. Protéine SdbA de Clostridium thermocellum, dont la séquence en acides aminés est la séquence complète de 631 acides aminés substantiellement telle que représentée sur l'IDS n° 1.

13. Fragment d'une protéine selon l'une des revendications 1 à 12 ou d'une protéine homologue, caractérisé en ce qu'il s'agit d'un domaine cohésine de type II.

15 14. Composé selon l'une des revendications 1 à 11, caractérisé en ce qu'il comporte au moins un fragment non protéique.

15. Fragment d'ADN, caractérisé en ce qu'il comporte au moins une séquence codant pour un domaine cohésine de type II.

16. Fragment d'ADN selon la revendication 15, codant pour la protéine SdbA ou fragment de celle-ci.

20 17. Fragment d'ADN, selon la revendication 15, caractérisé en ce qu'il comporte pour séquence substantiellement les nucléotides 82 à 573 dans l'IDS n° 1 codant pour le domaine cohésine de type II de SdbA.

25 18. Fragment d'ADN selon la revendication 15, comportant substantiellement la séquence de nucléotides 1 à 1893 de l'IDS n° 1 codant pour la protéine SdbA.

30 19. Fragment d'ADN selon la revendication 15, caractérisé en ce qu'il a substantiellement pour séquence l'une des séquences codant pour un domaine cohésine de OlpB choisies parmi la séquence des nucléotides 85 à 570, la séquence des nucléotides 619 à 1095 et la séquence des nucléotides 1225 à 1689 et la séquence des nucléotides 1819 à 2189 dans l'IDS n° 2.

20. Fragment d'ADN selon la revendication 15, caractérisé en ce qu'il a substantiellement pour séquence l'une des séquences codant pour un domaine cohésine de ORF2 choisies parmi la séquence des nucléotides 109 à 582 et la séquence des nucléotides n° 625 à 1092 dans l'IDS n° 3.

35 21. Fragment d'ADN caractérisé en ce qu'il a pour séquence une séquence complémentaire ou homologue ou complémentaire de l'homologue d'un fragment selon l'une des revendications 15 à 20.

22. Fragment d'ADN caractérisé en ce qu'il est capable de s'hybrider dans des conditions faiblement stringentes avec un fragment selon l'une des revendications 15 à 21.

23. Souche de Ecoli déposée à la CNCM de l'Institut Pasteur sous le n° I-1683 transformée par le plasmide pCT 1801.

24. Souche de Ecoli déposée à la CNCM de l'Institut Pasteur sous le n° I-1684 transformée par le plasmide pCT 1830.

25. Composé caractérisé en ce qu'il comporte au moins un domaine dockerine de type II.

26. Complexe comportant au moins un composé selon l'une des revendications 1 à 14 lié par une interaction C/D de type II avec un composé comportant au moins un domaine dockerine de type II, chaque composé constituant un élément du complexe.

27. Complexe selon la revendication 26, caractérisé en ce que l'affinité du complexe est au moins égal à 10<sup>5</sup> M.

28. Complexe selon l'une des revendications 26 et 27, caractérisé en ce qu'il comporte au moins trois éléments dont deux sont liés par une interaction C/D autre que de type II.

29. Complexe selon la revendication 28, caractérisé en ce que deux éléments sont liés par une interaction C/D de type I.

30. Complexe multimérique selon les revendications 28 et 29 caractérisé en ce qu'il comprend entre 1 et 50 éléments associés entre eux et de préférence 1 et 20.

31. Complexe selon la revendication 30 caractérisé en ce qu'il comprend au moins deux domaines d'interaction C/D de type II.

32. Complexe selon la revendication 30 caractérisé en ce qu'il comprend au moins une interaction C/D de type I associé à une interaction C/D de type II.

33. Complexe multimérique selon l'une des revendications 28 à 32, caractérisé en ce que les éléments du complexe sont essentiellement des protéines.

34. Complexe selon l'une des revendications 26 à 33, caractérisé en ce qu'au moins l'un des éléments comprend un fragment protéique riche en proline et/ou en hydroxy amino acide.

35. Complexe multimérique selon l'une des revendications 33 et 34, caractérisé en ce que certains des éléments du complexe sont des enzymes.

36. Vecteur d'expression comprenant un fragment d'ADN selon l'une des revendications 15 à 22, placé sous le contrôle d'éléments assurant son expression dans une cellule hôte.

5 37. Souche de E.coli transformée par un vecteur selon la revendication 36.

38. Procédé de préparation d'un polypeptide selon l'une des revendications 1 à 14, caractérisé en ce qu'on réalise la culture de cellules hôtes transformées à l'aide d'un vecteur selon la revendication 36 ou par culture d'une souche selon la revendication 37.

10 39. Composition enzymatique comprenant un complexe selon l'une des revendications 26 à 35.

40. Composition enzymatique selon la revendication 39, comprenant deux enzymes liées par une interaction C/D de type II.

15 41. Composition selon les revendications 39, caractérisée en ce que le complexe multimérique comporte un composé selon l'une des revendications 1 à 11 liée à un domaine dockerine de la protéine CipA, lié à une première enzyme, et le second composé comprenant un domaine dockerine d'une sous unité catalytique du complexe cellulytique de Clostridium thermocellum lié à une seconde enzyme.

20 42. Utilisation du complexe multimérique selon l'une des revendications 26 à 35, caractérisée en ce que ledit complexe multimérique potentialise la synergie des éléments du complexe.

25 43. Utilisation du complexe multimérique selon l'une des revendications 39 à 42, caractérisée en ce que ledit complexe assure la potentialisation de la composition enzymatique.

30 44. Procédé de détection d'un antigène ou d'un anticorps, caractérisé par la mise en contact d'un complexe multimérique selon l'une des revendications 26 à 35 avec une solution contenant un anticorps ou un antigène d'intérêt et la révélation de la réaction entre le complexe multimérique et l'antigène ou l'anticorps.

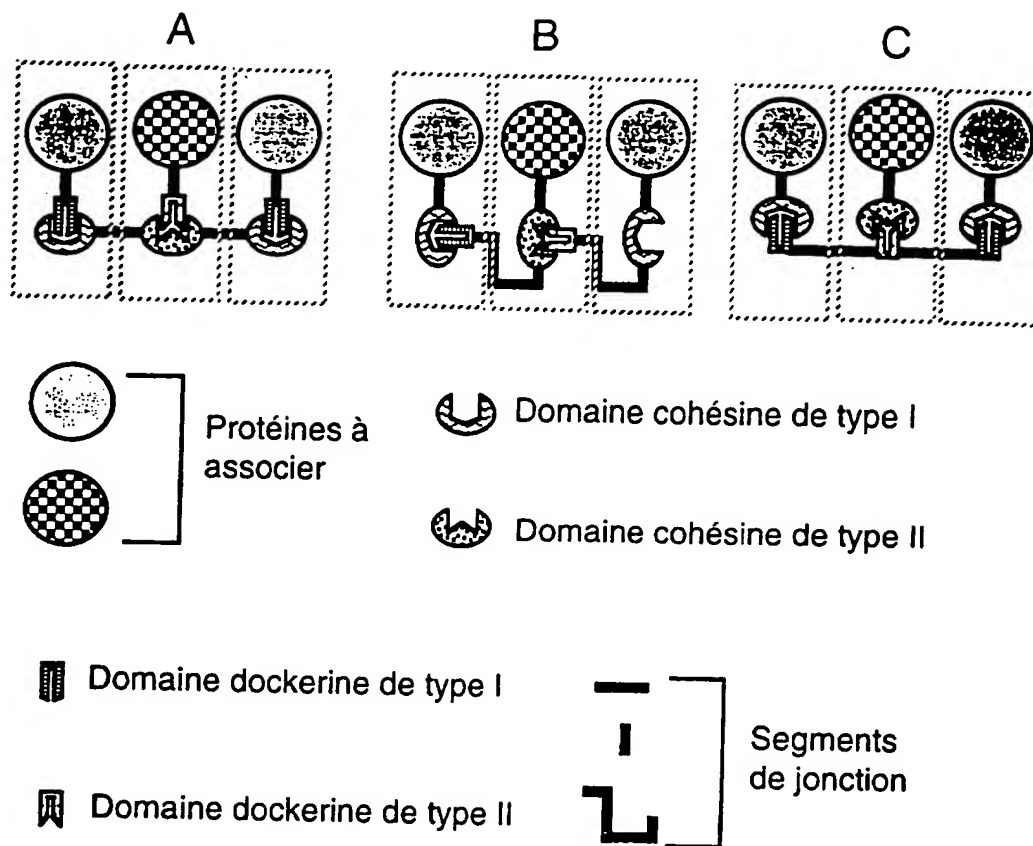


Figure 1

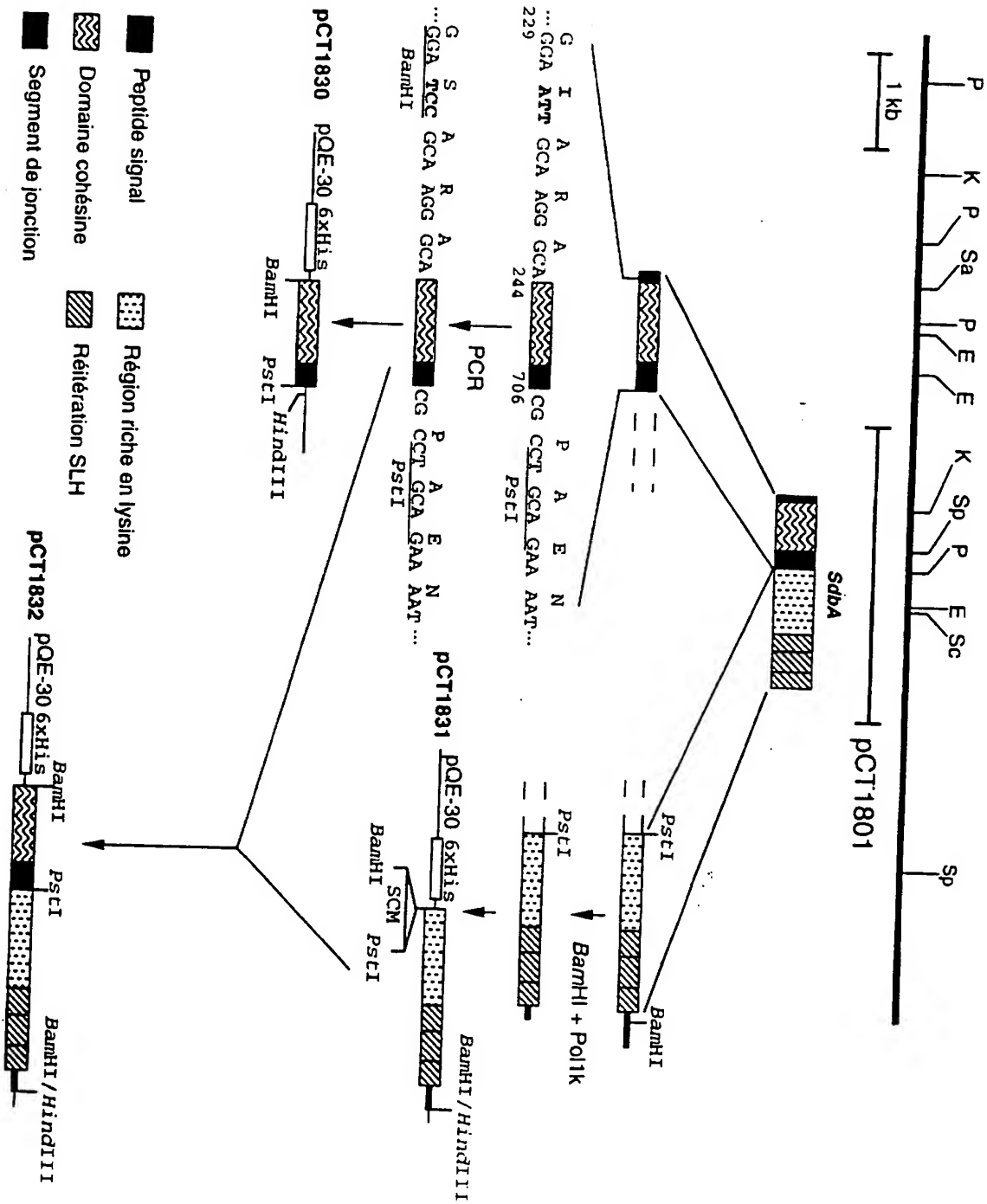


Figure 2

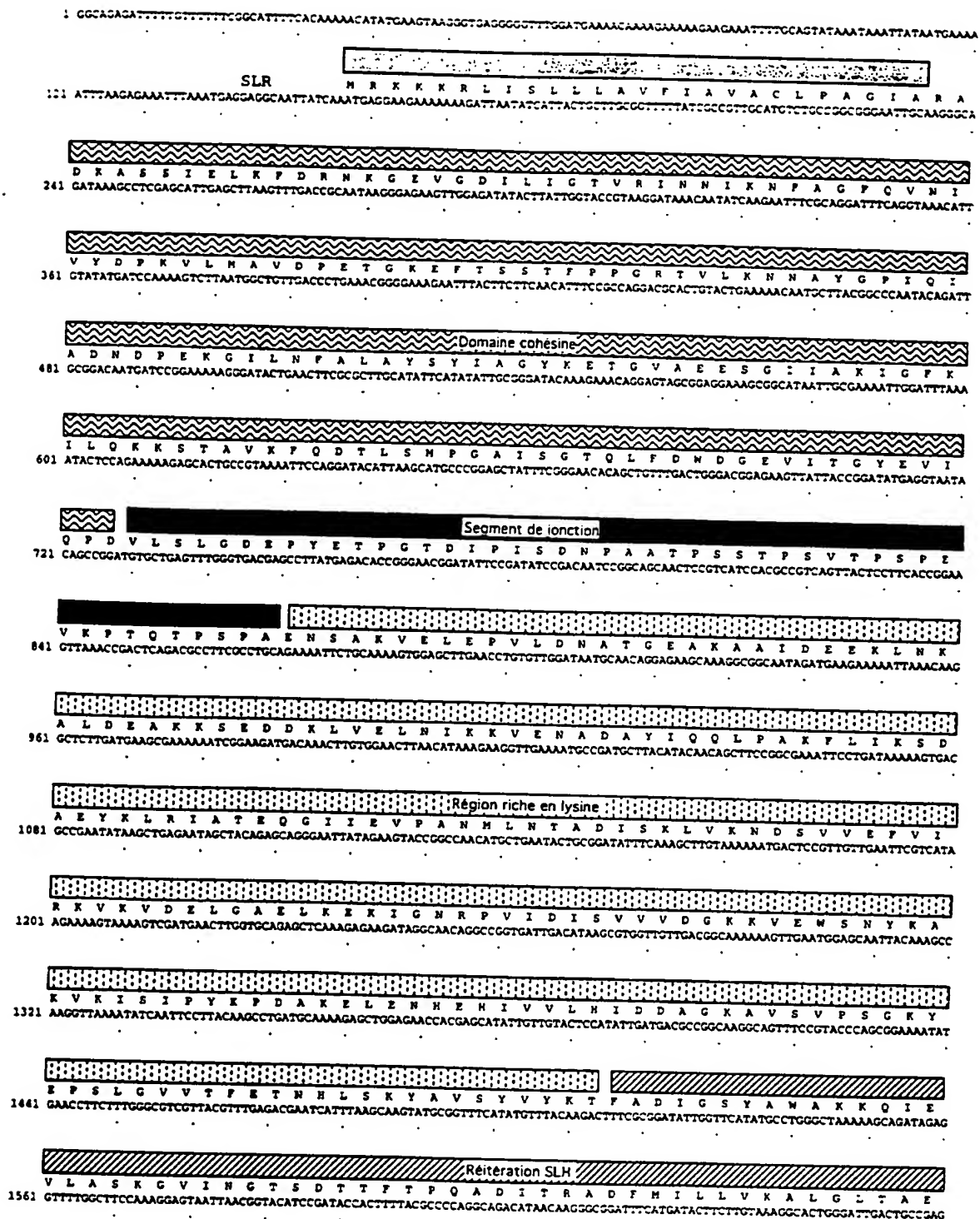


Figure 3

Reiteration SLH  
V T S N F D D V S E K D Y Y Y E Y V G I A K E L G I T T G V G N N K F N P K A K  
1681 GTTACTCCCAATTTTGATGATGTGTCCGAAAAAGACTACTATTATGAATACGTGGGAATTCGAAAAGAGCTTGAATTACGACAGGAGTGGAAACAACAAGTTCATCCGAAAGCCAAA

I T R Q D M H V L T T N A L R I A G K I S S T G T R A D V E R F S D K D Q I A S  
1801 ATTACAAGACAGGATATGATGGTACTTACAACAAATGCTCTCAGGATTGCAGGAAAAATATCGAGCACAGGAACCCGCGCTGATGTTGAAAGATTTCCGACAGGACCAGATAGCTTCA

Reiteration SLH  
Y A V E G V A T L V K E C I V V G S C D I I N P R C N A S R A E L A A I I Y K I  
1921 TATGCGGTGAAGGCGTTCCAACTTGTAAAAGAGGTATTGTAGTGGGAAGCGCGATATTATAATCCAAGGGGAAATGCTTCAAGAGCGGAAGTTCGACCAATCATATACAAGATT

Y Y K  
2041 TACTACAAGTAAATTTGTTTTGCATAAGTCAAGTGAAGGATAAACAGGGGATACGGCCAGGGTGAAAAGCCTTTGATGGGTGCTTCCCTAAAATTATATTCTGGTAAATATT

Figure 3 (suite)



27 RADKASSIELKFDRNKGEVGDILIGTVRINNIKNFAGFQVNIIVYDPKVLMAVDPET SbpA  
 28 .AEATPSIEMVLDKTEVHVGDVITATIKVNNIRKLAGYQLNIKFDPVLPVDPAT OlpB  
 207 .....LELDKTKVKVGDIIITATIKIENMKNFAGYQLNIKYDPTMLEAIELET OlpB  
 409 .....MELDKTKVKVGDIIITATIKIENMKNFAGYQLNIKYDPTMLEAIELET OlpB  
 607 .....MELDKTKVKEGDVIIATIRVNNIKNLAGYQIGIKYDPKYLEAFNIET OlpB  
 38 .....MELDKTKANIGDIIITATIRIDNINNFSGYQLNIKYDESYLOAVNPLT ORF2p  
 209 .....ALELDKTKVKVGDVIVATVKAKNMTSMAGIQVNIKYDPEVLQAIIDPAT ORF2p  
 83 GKEETSSTFPP..GRTVLKNNAYGPIQIADNDPEKGILNFAALAYSIIAGYKETGVA SbpA  
 84 GEEETDKSMPV..NRVLLTNSKYGPTPVAGNDIKSGIINFATGYNNLTAYKSSGID OlpB  
 254 GSATAKRTWEPVTGGTV..EQSDNYGKTTAVANDVGAGIINFAEAYSNLTKYRETGVA OlpB  
 456 GSATAKRTWEPVTGGTV..EQSDNYGKTTAVANDVGAGIINFAEAYSNLTKYRETGVA OlpB  
 654 GDPIDEGTWPAVGGTI..LKNRDYLP TGVA INNVSKGILNFAAYVYFDDYREEGKS OlpB  
 85 GEPIKKRTMPAVNGTVLEKGDQYSITEVVENNVDGILNFGKGYANLTEYRKSGKP ORF2p  
 257 GKPE TKETELV..DPELLSNREYNPLLTAVNDINSGIINYASCYVYWD SYRESGVS ORF2p  
 137 EESGIIAKIGFKILQKKSTAVKFQDTLSMPGAISGTQLFDWDGEVITGYEVIQIP SbpA  
 138 EHTGPIGEIGFKVLKKQNTSIRFEDTLSMPGAISGTSLFDWDAETITGYEVIQIP OlpB  
 309 EETGIIIGKIGERVVKAGSTAIRFEDTTAMPGAIEGTYMEDWYGENIKGYSVVQP OlpB  
 511 EETGIIIGKIGERVVKAGSTAIRFEDTTAMPGAIEGTYMEDWYGENIKGYSVVQP OlpB  
 709 EDTGIIIGKIGERVVKAEDETTIRFEELESMPGSIDGTYMFDWYLNRTSGYVVIQIP OlpB  
 141 ETTGIIIGKIGFKALKLGKTEIKFENTPVMPGAKEGTLLEFDWDAETITEYNVIQIP ORF2p  
 311 ESTGIIIGKVGEKVLKAANTVKLEETRFTPNSIDGTLVIDWYGOQIVGYKVIQIP ORF2p

Figure 4

098 EKAKQALEDQRK M1  
 264 EKLNKALDEAKK SbpA  
 278 EKLNKELEEGKK M9  
 289 EKLNKELEESKK PAM  
 450 EKLNKDLEESKK M12

Figure 5

1453 ..... AVIRGY... PDGSEPERNITTRAEAVITFAKLJGADESYGQASASP... OlpB  
 1496 YSDIAD..THWAAWAKPFATSOGLFKGY... PDGTHKPDQNTITRAEEAVITLHFTKVKQGEIMSKLATIDISNPK... OlpB  
 1567 FDDCVG..HWAQEFUEKLTSLGVISGY... PDGTFKQONVKKHSESVAGLHNHAFERGPLNGAP.KL..... OlpB  
 1627 FPDVNE..SYWAFGDVMDGALD..... PFIKGY... PGGLVQENNTITRAEAVITFAKLJGADENSAGKNSSIT... OlpB  
 241 ..... PFIKGY... PGGLVQENNTITRAEAVITFAKLJGADENSAGKNSSIT... OlpA  
 285 FKDLKD..SHWAAWAKFYVTEQNIFGGY... PDGTFMEPKSLTITRAEEAVITLHFTKVKQGEIMSKLATIDISNPK... OlpA  
 350 LKBJEG..HWAQYFETLVAKGVFKGY... PDGTFMEPKSLTITRAEEAVITLHFTKVKQGEIMSKLATIDISNPK... OlpA  
 410 FTDVVPV..NYWAYKDIAGGVIVY... PDGTFMEPKSLTITRAEEAVITLHFTKVKQGEIMSKLATIDISNPK... OlpA  
 454 FADYGS..YAWAKKQIEVLASKGVING... TSDTPTTPOADLITRADFMIDIVKAKGLTAEVTSN..... SbpA  
 513 FDDVSE..KDYVYEVYGIKELGIVTC... VGNKKNFKAKLITRODMVITTTNAIRIAGKISSGTGRADVER..... SbpA  
 580 FSDKDQIASVAVEGVATLVKEGIVVG... SGDIUNPRGNASTAEHNAITVYIYYK... SbpA  
 1682 FNDIKD..NMAKDVIEVLASRRHIVEG... MTDTOYEBNKTITRAEETAMILRLJLNKIDETYSGE..... Pul  
 1741 FSDVKS..GDWYANAEAAAYKACIEG... DGKNAIRNDSTITREETAMILRLJLNKIDETYSGE..... Pul  
 1805 FSDDKSISDWARNVANAALIGIVNG... EPNVFAFKGNATITAEAKATYGLLEKSGNI..... Pul  
 36 LNDFNKISGYAKEAVQSLVDAGVIOG... DANGNFNPLKTSRAEATITETNAIELEAEGDVN..... BspH  
 93 FKDVKA..DAWYYDAEATVENGIIIEG... VSATEFAPNKQLTTRSEAKLIVDAFELEGEGLSE..... BspH  
 153 FADASTVKPWAKEYUEIAVANGVINGSEANGKTNINAPITHOQFAWESHTYENVDATPKVDKIE..... BspH

Figure 6

DOCUMENTS CONSIDERES COMME PERTINENTS		Revendications concernées de la demande examinée
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	
D,Y	J. BACTERIOL. (1996), 178(4), 1200-3 CODEN: JOBAAY;ISSN: 0021-9193, XP000616410 LYTLE, BETSY ET AL: "Interactions of the CelS binding ligand with various receptor domains of the Clostridium thermocellum cellulosomal scaffolding protein, CipA" * page 1200, colonne 1 *	1,2
D,Y	J. BACTERIOL. (1994), 176(10), 2822-7 CODEN: JOBAAY;ISSN: 0021-9193, XP000616203 SALAMITOU, SYLVIE ET AL: "Recognition specificity of the duplicated segments present in Clostridium thermocellum endoglucanase CelD and in the cellulosome - integrating protein CipA" * page 2822, colonne 1 *	1,2
D,A	J. BACTERIOL. (1995), 177(9), 2451-9 CODEN: JOBAAY;ISSN: 0021-9193, XP000616404 LEMAIRE, MARC ET AL: "OlpB, a new outer layer protein of Clostridium thermocellum, and binding of its S-layer-like domains to components of the cell envelope" * le document en entier *	1
D,A	J. BACTERIOL. (1993), 175(7), 1891-9 CODEN: JOBAAY;ISSN: 0021-9193, XP000616403 FUJINO, TSUCHIYOSHI ET AL: "Organization of a Clostridium thermocellum gene cluster encoding the cellulosomal scaffolding protein CipA and a protein possibly involved in attachment of the cellulosome to the cell surface" * page 1897 *	1
---		
-/--		
Date d'achèvement de la recherche		Examineur
28 Janvier 1997		Delanghe, L
<p><b>CATEGORIE DES DOCUMENTS CITES</b></p> <p>X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : pertinent à l'encontre d'au moins une revendication ou arrière-plan technologique général O : divulgation non-écrite P : document intercalaire</p> <p>T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons</p> <p>--- &amp; : membre de la même famille, document correspondant</p>		

DOCUMENTS CONSIDERES COMME PERTINENTS		Revendications concernées de la demande examinée
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	
D,Y	J. BACTERIOL. (1996), 178(4), 1200-3 CODEN: JOBAAY;ISSN: 0021-9193, XP000616410 LYTLE, BETSY ET AL: "Interactions of the Cels binding ligand with various receptor domains of the Clostridium thermocellum cellulosomal scaffolding protein, CipA" * page 1200, colonne 1 *	1,2
D,Y	--- J. BACTERIOL. (1994), 176(10), 2822-7 CODEN: JOBAAY;ISSN: 0021-9193, XP000616203 SALAMITOU, SYLVIE ET AL: "Recognition specificity of the duplicated segments present in Clostridium thermocellum endoglucanase CelD and in the cellulosome - integrating protein CipA" * page 2822, colonne 1 *	1,2
D,A	--- J. BACTERIOL. (1995), 177(9), 2451-9 CODEN: JOBAAY;ISSN: 0021-9193, XP000616404 LEMAIRE, MARC ET AL: "OlpB, a new outer layer protein of Clostridium thermocellum, and binding of its S-layer-like domains to components of the cell envelope" * le document en entier *	1
D,A	--- J. BACTERIOL. (1993), 175(7), 1891-9 CODEN: JOBAAY;ISSN: 0021-9193, XP000616403 FUJINO, TSUCHIYOSHI ET AL: "Organization of a Clostridium thermocellum gene cluster encoding the cellulosomal scaffolding protein CipA and a protein possibly involved in attachment of the cellulosome to the cell surface" * page 1897 *	1
--- -/--		
		DOMAINES TECHNIQUES RECHERCHES (Int.CL.6)
		C07K C12N
Date d'achèvement de la recherche		Examineur
28 Janvier 1997		Delanghe, L
CATEGORIE DES DOCUMENTS CITES		
X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : pertinent à l'encontre d'au moins une revendication ou arrière-plan technologique général O : divulgation non-écrite P : document intercalaire		T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant